PATENT ABSTRACTS OF JAPAN

(11)Publication number : 11-296552
(43)Date of publication of application : 29.10.1999

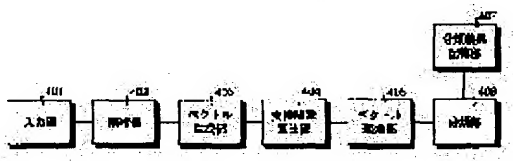
(51)Int.CI. G06F 17/30
G06F 17/27

(21)Application number : 10-115907
(22)Date of filing : 13.04.1998

(71)Applicant : RICOH CO LTD
(72)Inventor : KENMOCHI EIJI
MIYAJI TATSUO
SHIMADA ATSUO
TAKEYA KAZUHISA
NAKAJIMA AKIKO
NAGATSUKA TETSUO
YAMAZAKI MAKOTO
FUJITA KATSUHIKO

(54) DEVICE AND METHOD FOR CLASSIFYING DOCUMENT AND COMPUTER-READABLE RECORDING MEDIUM WHERE PROGRAM ALLOWING COMPUTER TO IMPLEMENT SAME METHOD IS RECORDED

(57)Abstract:
PROBLEM TO BE SOLVED: To classify documents according to the similarities between the documents repeatedly in a short time with good efficiency so that the intention of the operator is reflected.
SOLUTION: This device is equipped with an input part 401 which inputs document data, an analysis part 402 which obtains analytic information by analyzing the inputted document data, a vector generation part 403 which generates document feature vectors for the document data according to the obtained analytic information, a conversion function calculation part 404 which calculates a representation space conversion function for projecting the generated document feature vectors in a space wherein the similarities between the document feature vectors are reflected, a vector conversion part 405 which converts the document feature vectors generated by the vector generation part 403 by using the calculated representation space converting function, a classification part 406 which classifies documents according to the similarities between the converted document feature vectors, and a classification result storage part 407 which stores the results of the classified documents.



LEGAL STATUS

[Date of request for examination] 18.09.2002

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12)**公開特許公報（Ａ）**　(11)特許出願公開番号

特開平11－296552

(43)公開日　平成11年(1999)10月29日

(51)Int.Cl.[6]	識別記号	ＦＩ		
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ　15/401		３１０Ｄ
17/27			15/20	５５０Ｅ
			15/40	３７０Ａ

審査請求　未請求　請求項の数33　ＦＤ　（全 24 頁）

(21)出願番号　特願平10－115907

(22)出願日　平成10年(1998)４月13日

(71)出願人　000006747
株式会社リコー
東京都大田区中馬込１丁目３番６号

(72)発明者　剣持　栄治
東京都大田区中馬込１丁目３番６号　株式会社リコー内

(72)発明者　宮地　達生
東京都大田区中馬込１丁目３番６号　株式会社リコー内

(72)発明者　嶋田　敦夫
東京都大田区中馬込１丁目３番６号　株式会社リコー内

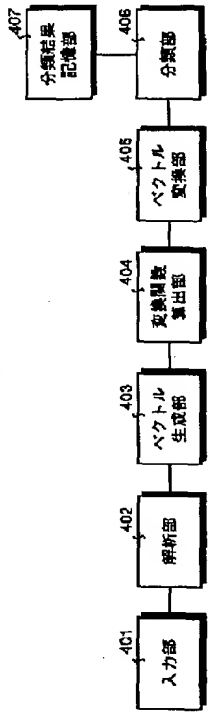
最終頁に続く

(54)【発明の名称】　文書分類装置、文書分類方法およびその方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体

(57)【要約】

【課題】　文書間の類似性に基づいて文書分類をおこなう際、操作者の意図を反映する文書分類を短時間で効率良く繰り返しをおこなうことを課題とする。

【解決手段】　文書データを入力する入力部４０１と、入力された文書データを解析し解析情報を得る解析部４０２と、得られた解析情報に基づいて文書データに対する文書特徴ベクトルを生成するベクトル生成部４０３と、生成された文書特徴ベクトルが文書特徴ベクトル間の類似性を反映する空間に射影されるための表現空間変換関数を算出する変換関数算出部４０４と、算出された表現空間変換関数をもちいてベクトル生成部４０３により生成された文書特徴ベクトルを変換するベクトル変換部４０５と、変換された文書特徴ベクトル間の類似度に基づいて文書を分類する分類部４０６と、分類された文書分類の結果を記憶する分類結果記憶部４０７とを備える。



1

【特許請求の範囲】

【請求項１】　文書データを入力する入力手段と、
前記入力手段により入力された文書データを解析し解析情報を得る解析手段と、
前記解析手段により得られた解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成するベクトル生成手段と、
前記ベクトル生成手段により生成された文書特徴ベクトルが文書特徴ベクトル間の類似性を反映する空間に射影されるための表現空間変換関数を算出する変換関数算出手段と、
前記変換関数算出手段により算出された表現空間変換関数をもちいて前記ベクトル生成手段により生成された文書特徴ベクトルを変換するベクトル変換手段と、
前記ベクトル変換手段により変換された文書特徴ベクトル間の類似度に基づいて文書を分類する分類手段と、
前記分類手段により分類された文書分類の結果を記憶する分類結果記憶手段と、
を備えたことを特徴とする文書分類装置。

【請求項２】　前記ベクトル生成手段により生成された文書特徴ベクトル間の内積を算出する内積算出手段を備え、
前記変換関数算出手段は、前記内積算出手段により算出された内積をもちいて表現空間変換関数を算出することを特徴とする請求項１に記載の文書分類装置。

【請求項３】　前記入力手段により入力された文書の作成者、作成日等の文書データの文書間類似情報を設定する文書間類似情報設定手段を備え、
前記変換関数算出手段は、前記内積算出手段により算出された内積および前記文書間類似情報設定手段により設定された文書間類似情報をもちいて表現空間変換関数を算出することを特徴とする請求項２に記載の文書分類装置。

【請求項４】　さらに、前記ベクトル生成手段により生成された文書特徴ベクトルを記憶するベクトル記憶手段と、
前記変換関数算出手段により算出された表現空間変換関数を記憶する変換関数記憶手段と、
を備えたことを特徴とする請求項１～３のいずれか一つに記載された文書分類装置。

【請求項５】　さらに、前記ベクトル変換手段により文書特徴ベクトルを変更する前に、前記解析手段により抽出される単語が有する特性により構成される規則をもちいて前記文書特徴ベクトルおよび／または文書特徴ベクトルを構成する特徴次元を操作することにより前記文書特徴ベクトルを修正するベクトル修正手段を備えたことを特徴とする請求項１～４のいずれか一つに記載された文書分類装置。

【請求項６】　前記ベクトル修正手段において文書特徴ベクトルを修正することにより特徴次元が変更された場合に、前記変更された特徴次元により前記ベクトル変換手段において前記文書特徴ベクトルが適切に変換できるように、前記変換関数算出手段により算出された表現空間変換関数を修正する変換関数修正手段を備えたことを特徴とする請求項５に記載の文書分類装置。

【請求項７】　さらに、前記表現空間変換関数の特徴次元の操作に関する指示をする変換関数修正指示手段と、
前記変換関数修正指示手段により指示された特徴次元の操作に関する指示内容に基づいて、前記表現空間変換関数を修正する変換関数修正手段と、を備えたことを特徴とする請求項１～５のいずれか一つに記載の文書分類装置。

【請求項８】　前記変換関数修正指示手段により指示された特徴次元の操作に関する指示内容が、任意の文書ベクトルデータをもちいて前記表現空間変換関数の特徴次元を操作するものであることを特徴とする請求項７に記載の文書分類装置。

【請求項９】　前記変換関数修正指示手段により指示された特徴次元の操作に関する指示内容が、文書特徴ベクトルをもちいて前記表現空間変換関数の特徴次元を操作するものであることを特徴とする請求項７に記載の文書分類装置。

【請求項１０】　前記変換関数修正指示手段により指示された特徴次元の操作に関する指示内容が、前記解析手段により得られた解析情報をもちいて前記表現空間変換関数の特徴次元を操作するものであることを特徴とする請求項７に記載の文書分類装置。

【請求項１１】　前記変換関数修正指示手段により指示された特徴次元の操作に関する指示内容が、前記分類結果記憶手段により記憶された分類結果をもちいて前記表現空間変換関数の特徴次元を操作するものであることを特徴とする請求項７に記載の文書分類装置。

【請求項１２】　初期クラスタ重心を指定する初期重心指定手段と、
前記初期重心指定手段により指定された初期クラスタ重心を登録する初期重心登録手段とを備え、
前記分類手段は、前記初期重心登録手段により登録された初期クラスタ重心にしたがって文書を分類することを特徴とする請求項１～１１のいずれか一つに記載の文書分類装置。

【請求項１３】　前記初期重心指定手段により指定される初期クラスタ重心として任意の文書ベクトルデータを指定することを特徴とする請求項１２に記載の文書分類装置。

【請求項１４】　前記初期重心指定手段により指定される初期クラスタ重心として文書特徴ベクトルを指定することを特徴とする請求項１２に記載の文書分類装置。

【請求項１５】　前記初期重心指定手段により指定される初期クラスタ重心として前記解析手段により得られた解析情報を指定することを特徴とする請求項１２に記載

の文書分類装置。

【請求項１６】　前記初期重心指定手段により指定される初期クラスタ重心として前記分類結果記憶手段により記憶された分類結果を指定することを特徴とする請求項１２に記載の文書分類装置。

【請求項１７】　文書データを入力する第１工程と、
前記第１工程により入力された文書データを解析し解析情報を得る第２工程と、
前記第２工程により得られた解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成する第３工程と、
前記第３工程により生成された文書特徴ベクトルが文書特徴ベクトル間の類似性を反映する空間に射影されるための表現空間変換関数を算出する第４工程と、
前記第４工程により算出された表現空間変換関数をもちいて前記第３工程により生成された文書特徴ベクトルを変換する第５工程と、
前記第５工程により変換された文書特徴ベクトル間の類似度に基づいて文書を分類する第６工程と、
前記第６工程分類手段により分類された文書分類の結果を記憶する第７工程と、
を含んだことを特徴とする文書分類方法。

【請求項１８】　前記第３工程により生成された文書特徴ベクトル間の内積を算出する第８工程を含み、
前記第４工程は、前記第８工程により算出された内積をもちいて表現空間変換関数を算出することを特徴とする請求項１７記載の文書分類方法。

【請求項１９】　前記第１工程により入力された文書の作成者、作成日等の文書データの文書間類似情報を設定する第９工程を含み、
前記第４工程は、前記第８工程により算出された内積および前記第９工程により設定された文書間類似情報をもちいて表現空間変換関数を算出することを特徴とする請求項１８に記載の文書分類方法。

【請求項２０】　さらに、前記第３工程により生成された文書特徴ベクトルを記憶する第１０工程と、
前記第４工程により算出された表現空間変換関数を記憶する第１１工程と、
を含んだことを特徴とする請求項１７～１９のいずれか一つに記載された文書分類方法。

【請求項２１】　さらに、前記第５工程により文書特徴ベクトルを変更する前に、前記第２工程により抽出される単語が有する特性により構成される規則をもちいて前記文書特徴ベクトルおよび／または文書特徴ベクトルを構成する特徴次元を操作することにより前記文書特徴ベクトルを修正する第１２工程を含んだことを特徴とする請求項１７～２０のいずれか一つに記載された文書分類方法。

【請求項２２】　前記第１２工程において文書特徴ベクトルを修正することにより特徴次元が変更された場合に、前記変更された特徴次元により第５工程において前記文書特徴ベクトルが適切に変換できるように、前記第４工程により算出された表現空間変換関数を修正する第１３工程を含んだことを特徴とする請求項２１に記載の文書分類方法。

【請求項２３】　さらに、前記表現空間変換関数の特徴次元の操作に関する指示をする第１４工程と、
前記第１４工程により指示された特徴次元の操作に関する指示内容に基づいて、前記表現空間変換関数を修正する第１５工程と、
を含んだことを特徴とする請求項１７～２１のいずれか一つに記載の文書分類方法。

【請求項２４】　前記第１５工程により指示された特徴次元の操作に関する指示内容が、任意の文書ベクトルデータをもちいて前記表現空間変換関数の特徴次元を操作するものであることを特徴とする請求項２３に記載の文書分類方法。

【請求項２５】　前記第１５工程により指示された特徴次元の操作に関する指示内容が、文書特徴ベクトルをもちいて前記表現空間変換関数の特徴次元を操作するものであることを特徴とする請求項２３に記載の文書分類方法。

【請求項２６】　前記第１５工程により指示された特徴次元の操作に関する指示内容が、前記第２工程により得られた解析情報をもちいて前記表現空間変換関数の特徴次元を操作するものであることを特徴とする請求項２３に記載の文書分類方法。

【請求項２７】　前記第１５工程により指示された特徴次元の操作に関する指示内容が、前記第７工程により記憶された分類結果をもちいて前記表現空間変換関数の特徴次元を操作するものであることを特徴とする請求項２３に記載の文書分類方法。

【請求項２８】　初期クラスタ重心を指定する第１６工程と、
前記第１６工程により指定された初期クラスタ重心を登録する第１７工程とを含み、
前記第６工程は、前記第１７工程により登録された初期クラスタ重心にしたがって文書を分類することを特徴とする請求項１７～２７のいずれか一つに記載の文書分類方法。

【請求項２９】　前記第１６工程により指定される初期クラスタ重心として任意の文書ベクトルデータを指定することを特徴とする請求項２８に記載の文書分類方法。

【請求項３０】　前記第１６工程により指定される初期クラスタ重心として文書特徴ベクトルを指定することを特徴とする請求項２８に記載の文書分類方法。

【請求項３１】　前記第１６工程により指定される初期クラスタ重心として前記第２工程により得られた解析情報を指定することを特徴とする請求項２８に記載の文書分類方法。

【請求項３２】　前記第１６工程により指定される初期クラスタ重心として前記第７工程により記憶された分類結果を指定することを特徴とする請求項２８に記載の文書分類方法。

【請求項３３】　前記請求項１７～３２のいずれか一つに記載された方法をコンピュータに実行させるプログラムを記録したことを特徴とするコンピュータ読み取り可能な記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】この発明は、文書間の類似性に基づいて文書を分類する文書分類装置、文書分類方法およびその方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体に関する。

【０００２】

【従来の技術】近年インターネット等の普及により大量の文書情報へのアクセスが可能になり、収集した大量の文書情報を意味のあるカテゴリーに分類し、文書群の構造を把握するなどの知的作業がおこなわれるようになっていきている。大量の文書情報を操作者が手動で分類する場合、人的／時間的コストが膨大なものになり、また分類をおこなう際にもちいる知識を分類をする操作者のみが有することになるため、分類をおこなう担当の操作者が変わると分類基準も変わってしまうことになる。

【０００３】したがって、文書群をいかに人間が分類をおこなうような分類基準によって自動的に分類することができるかが重要な課題となる。すなわち、意味的に類似している文書は同一のカテゴリーに分類され、また、分類をする工程において生成される各分類カテゴリーは操作者が文類実行前に意図しているような分類カテゴリーとなるように構成された文書分類装置の出現が望まれている。

【０００４】文書の自動分類装置の従来技術としては、たとえば特開平７－３６８９７号公報に記載されているように、文書を単語を特徴とする文書ベクトルとみなし、クラスタリング手法をもちいてこれらの文書ベクトルを群分けし、群分けした文書ベクトルに基づいて文書の自動分類をおこなうものがある。

【０００５】

【発明が解決しようとする課題】しかしながら、上記従来技術の文書分類装置は、分類対象文書に含まれる単語を特徴量とする文書特徴ベクトルをもちいて、その文書特徴ベクトルに対しクラスタリング手法を適用して分類をおこなうため、単語の多義性／同義性により文書の意味的な関連性を反映した分類結果を得ることが困難となるという問題があった。

【０００６】この単語の多義性／同義性の問題を解決するものとしては、米国特許第４８３９８５３号公報に記載されているように、文書間の内積行列に特異値分解を適用するものがある。すなわち、文書間の単語の共起性をもとに生成される潜在的意味空間といわれる空間へ、文書と単語を射影することにより意味的な関連性を反映した文書検索をおこなうものである。

【０００７】また、「Ｐｒｏｊｅｃｔｉｏｎｓ ｆｏｒ Ｅｆｆｉｃｉｅｎｔ Ｄｏｃｕｍｅｎｔ Ｃｌｕｓｔｅｒｉｎｇ（著者名：Ｈｉｎｒｉｃｈ Ｓｃｈｕｔｚｅ ａｎｄ Ｃｒａｉｎｇ Ｓｉｌｖｅｒｓｔｅｉｎ，学会名：ＡＣＭ，　論文名：Ｐｒｏｃｅｅｄｉｎｇｓ ｏｆ ＳＩＧＩＲ，　ページ：７４－８１，　発行年：１９９７）」においては、上記潜在的意味空間において文書分類を実施しているものがある。さらに、「Ｒｅｐｒｅｓｅｎｔａｔｉｎｇ ＤｏｃｕｍｅｎｔｓＵｓｉｎｇ ａｎ Ｅｘｐｌｉｃｉｔ Ｍｏｄｅｌ ｏｆ Ｔｈｅｉｒ Ｓｉｍｉｌａｒｉｔｉｅｓ（著者名：Ｂｒｉａｎ Ｔ．Ｂａｒｔｅｌｌ，Ｇａｒｒｉｓｏｎ Ｗ．Ｃｏｔｔｒｅｌｌ，ａｎｄ Ｒｉｃｈａｒｄ Ｋ．Ｂｅｌｅｗ，　論文名：Ｊｏｕｒｎａｌ ｏｆ ｔｈｅ Ａｍｅｒｉｃａｎ Ｓｏｃｉｅｔｙ ｆｏｒ Ｉｎｆｏｒｍａｔｉｏｎ Ｓｃｉｅｎｃｅ，　学会名：ｔｈｅ Ａｍｅｒｉｃａｎ Ｓｏｃｉｅｔｙ ｆｏｒ Ｉｎｆｏｒｍａｔｉｏｎ Ｓｃｉｅｎｃｅ，　ページ：２５４－２７１，Ｖｏｌ．４６ Ｎｏ．４，　発行年：１９９５）」においては、上記潜在的意味空間への変換手法を一般化し、文書間の内積行列に、文書が有する他の文書への参照情報から生成される共参照情報などを付加した行列をもちいて、これらの類似性を反映する空間へ文書や単語を射影するための表現空間変換関数を導出しているものがある。

【０００８】これらの従来技術の手法で生成される射影空間の各次元は複数の単語が意味的に結合した概念的なものであるが、どの特徴次元を使って文書分類あるいは文書検索をおこなうかは、特異値分解を適用する際に算出される特異値の大きさのみを基準として決定される。このため、分類実行時にもちいられる特徴次元の選択においては、操作者の意図が反映されることは困難であり、このため分類結果が操作者の意図する結果と異なってしまうという問題点があった。

【０００９】また、従来の他の文書分類方法では、文書の意味的な関連性を反映した文書分類をおこなうために、文書を変換するするための表現空間変換関数を算出する部分と実際に前記表現空間変換関数をもちいて変換された文書の文書分類をおこなう部分とを連続的に処理しているが、表現空間変換関数を算出する部分は非常に計算時間を費やす処理であるため、結果として一回の文書分類に要する時間も膨大なものになるという問題点があった。

【００１０】この発明は、上述した従来例による問題点を解消するため、操作者の意図を反映する文書分類を短時間で効率良く繰り返しをおこなうことができる文書分

類装置、文書分類方法およびその方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体を提供することを目的とする。

【００１１】

【課題を解決するための手段】上述した課題を解決し、目的を達成するため、請求項１の発明に係る文書分類装置は、文書データを入力する入力手段と、前記入力手段により入力された文書データを解析し解析情報を得る解析手段と、前記解析手段により得られた解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成するベクトル生成手段と、前記ベクトル生成手段により生成された文書特徴ベクトルが文書特徴ベクトル間の類似性を反映する空間に射影されるための表現空間変換関数を算出する変換関数算出手段と、前記変換関数算出手段により算出された表現空間変換関数をもちいて前記ベクトル生成手段により生成された文書特徴ベクトルを変換するベクトル変換手段と、前記ベクトル変換手段により変換された文書特徴ベクトル間の類似度に基づいて文書を分類する分類手段と、前記分類手段により分類された文書分類の結果を記憶する分類結果記憶手段と、を備えたことを特徴とする。

【００１２】この請求項１の発明によれば、分類対象である文書群での文書間の類似性に基づいて、各文書をそれら文書間の意味的な関連性を反映しうる表現空間へ変換するための表現空間変換関数を算出し、その表現空間で文書分類をおこなうことにより、操作者の意図を反映しうる文書分類を実現することが可能である。

【００１３】また、請求項２に係る文書分類装置は、請求項１の発明において、前記ベクトル生成手段により生成された文書特徴ベクトル間の内積を算出する内積算出手段を備え、前記変換関数算出手段が、前記内積算出手段により算出された内積をもちいて表現空間変換関数を算出することを特徴とする。

【００１４】この請求項２の発明によれば、表現空間変換関数を導出する際に必要となる文書間の類似性として文書特徴ベクトル間の内積をもちいることにより、文書間の意味的な関連性を反映した文書分類をおこなうことが可能である。

【００１５】また、請求項３に係る文書分類装置は、請求項２の発明において、前記入力手段により入力された文書の作成者、作成日等の文書データの文書間類似情報を設定する文書間類似情報設定手段を備え、前記変換関数算出手段が、前記内積算出手段により算出された内積および前記文書間類似情報設定手段により設定された文書間類似情報をもちいて表現空間変換関数を算出することを特徴とする。

【００１６】この請求項３の発明によれば、表現空間変換関数を導出する際に必要となる文書間の類似性として文書特徴ベクトル間の内積に加え、文書の作成者や作成日などの文書間類似情報ももちいることにより、文書間の意味的な関連性を反映した文書分類をおこなうことが可能である。

【００１７】また、請求項４に係る文書分類装置は、請求項１～３の発明において、さらに、前記ベクトル生成手段により生成された文書特徴ベクトルを記憶するベクトル記憶手段と、前記変換関数算出手段により算出された表現空間変換関数を記憶する変換関数記憶手段と、を備えたことを特徴とする。

【００１８】この請求項４の発明によれば、算出する文書特徴ベクトルと表現空間変換関数を記憶することにより、表現空間変換関数を算出する部分と実際に前記表現空間変換関数をもちいて変換された文書をもちいて文書分類をおこなう部分とを分離して処理するので、その都度、表現空間変換関数を算出することなしに文書分類を実行でき、さらに、前記文書特徴ベクトル変換部でもちいる表現空間変換関数として、事前に他の文書特徴ベクトルに基づいて生成された表現空間変換関数をもちいることもできるため、文書分類の繰り返し実行を短時間で効率良くおこなうことが可能である。

【００１９】また、請求項５に係る文書分類装置は、請求項１～４のいずれか一つの発明において、さらに、前記ベクトル変換手段により文書特徴ベクトルを変更する前に、前記解析手段により抽出される単語が有する特性により構成される規則をもちいて前記文書特徴ベクトルおよび／または文書特徴ベクトルを構成する特徴次元を操作することにより前記文書特徴ベクトルを修正するベクトル修正手段を備えたことを特徴とする。

【００２０】この請求項５の発明によれば、文書分類の繰り返し実行をおこなう際、個々の分類実行ごとに、文書特徴ベクトルやそれらを構成する特徴次元を操作することで、各分類ごとに異なる単語を削除して文書分類を実行する等の分類対象文書の範囲の変更や分類をおこなう空間の変更をおこなうことが可能である。

【００２１】また、請求項６に係る文書分類装置は、請求項５の発明において、前記ベクトル修正手段において文書特徴ベクトルを修正することにより特徴次元が変更された場合に、前記変更された特徴次元により前記ベクトル変換手段において前記文書特徴ベクトルが適切に変換できるように、前記変換関数算出手段により算出された表現空間変換関数を修正する変換関数修正手段を備えたことを特徴とする。

【００２２】この請求項６の発明によれば、表現空間変換関数が文書特徴ベクトルの内積をに基づいて算出される場合、表現空間変換関数をもちいて変換された文書をもちいて文書分類をおこなう部分において、文書特徴ベクトルやその特徴次元が操作された場合に生じる表現空間変換関数の不整合を簡便に修正することができるので、より適正な文書特徴ベクトルの変換をおこなうことが可能である。

【００２３】また、請求項７に係る文書分類装置は、請

求項１～５のいずれか一つの発明において、さらに、前記表現空間変換関数の特徴次元の操作に関する指示をする変換関数修正指示手段と、前記変換関数修正指示手段により指示された特徴次元の操作に関する指示内容に基づいて、前記表現空間変換関数を修正する変換関数修正手段と、を備えたことを特徴とする。

【００２４】この請求項７の発明によれば、前記表現空間変換関数をもちいて構成される空間の特徴次元について操作者が簡便な操作をすることにより、操作者の意図を反映しうる文書分類をおこなうことが可能である。

【００２５】また、請求項８に係る文書分類装置は、請求項７の発明において、前記変換関数修正指示手段により指示された特徴次元の操作に関する指示内容が、任意の文書ベクトルデータをもちいて前記表現空間変換関数の特徴次元を操作するものであることを特徴とする。

【００２６】この請求項８の発明によれば、前記表現空間変換関数をもちいて構成される空間の特徴次元について、操作者により指示された分類対象以外の任意の文書ベクトルデータをもちいての簡便な操作をすることにより、操作者の意図を反映しうる文書分類をおこなうことが可能である。

【００２７】また、請求項９に係る文書分類装置は、請求項７の発明において、前記変換関数修正指示手段により指示された特徴次元の操作に関する指示内容が、文書特徴ベクトルをもちいて前記表現空間変換関数の特徴次元を操作するものであることを特徴とする。

【００２８】この請求項９の発明によれば、前記表現空間変換関数をもちいて構成される空間の特徴次元について、操作者により指示された文書特徴ベクトルをもちいての簡便な操作をすることにより、操作者の意図を反映しうる文書分類をおこなうことが可能である。

【００２９】また、請求項１０に係る文書分類装置は、請求項７の発明において、前記変換関数修正指示手段により指示された特徴次元の操作に関する指示内容が、前記解析手段により得られた解析情報をもちいて前記表現空間変換関数の特徴次元を操作するものであることを特徴とする。

【００３０】この請求項１０の発明によれば、前記表現空間変換関数をもちいて構成される空間の特徴次元を、操作者により指示された解析情報をもちいての簡便な操作をすることにより、操作者の意図を反映しうる文書分類をおこなうことが可能である。

【００３１】また、請求項１１に係る文書分類装置は、請求項７の発明において、前記変換関数修正指示手段により指示された特徴次元の操作に関する指示内容が、前記分類結果記憶手段により記憶された分類結果をもちいて前記表現空間変換関数の特徴次元を操作するものであることを特徴とする。

【００３２】この請求項１１の発明によれば、前記表現空間変換関数をもちいて構成される空間の特徴次元を、操作者により指示された事前に分類された分類結果をもちいての簡便な操作をすることにより、操作者の意図を反映しうる文書分類をおこなうことが可能である。

【００３３】また、請求項１２に係る文書分類装置は、請求項１～１１のいずれか一つの発明において、初期クラスタ重心を指定する初期重心指定手段と、前記初期重心指定手段により指定された初期クラスタ重心を登録する初期重心登録手段とを備え、前記分類手段は、前記初期重心登録手段により登録された初期クラスタ重心にしたがって文書を分類することを特徴とする。

【００３４】この請求項１２の発明によれば、文書分類手法として、非階層型クラスタリング手法をもちいて、その際に必要となる初期クラスタ重心を、操作者が任意に指定することができ、その指定された初期クラスタ重心にしたがって文書分類をおこなうので、操作者の意図を反映する文書分類をおこなうことが可能である。

【００３５】また、請求項１３に係る文書分類装置は、請求項１２の発明において、前記初期重心指定手段により指定される初期クラスタ重心として任意の文書ベクトルデータを指定することを特徴とする。

【００３６】この請求項１３の発明によれば、文書分類手法として、非階層型クラスタリング手法をもちいて、その際に必要となる初期クラスタ重心として、分類対象以外の任意の文書をもちいることができるので、操作者の意図を反映する文書分類をおこなうことが可能である。

【００３７】また、請求項１４に係る文書分類装置は、請求項１２の発明において、前記初期重心指定手段により指定される初期クラスタ重心として文書特徴ベクトルを指定することを特徴とする。

【００３８】この請求項１４の発明によれば、文書分類手法として、非階層型クラスタリング手法をもちいて、その際に必要となる初期クラスタ重心として、文書特徴ベクトルをもちいることができるので、操作者の意図を反映する文書分類をおこなうことが可能である。

【００３９】また、請求項１５に係る文書分類装置は、請求項１２の発明において、前記初期重心指定手段により指定される初期クラスタ重心として前記解析手段により得られた解析情報を指定することを特徴とする。

【００４０】この請求項１５の発明によれば、文書分類手法として、非階層型クラスタリング手法をもちいて、その際に必要となる初期クラスタ重心として、分類対象文書を文書解析部に作用させた結果得られる単語等の解析情報をもちいることができるので、操作者の意図を反映する文書分類をおこなうことが可能である。

【００４１】また、請求項１６に係る文書分類装置は、請求項１２の発明において、前記初期重心指定手段により指定される初期クラスタ重心として前記分類結果記憶手段により記憶された分類結果を指定することを特徴とする。

【００４２】この請求項１６の発明によれば、文書分類手法として、非階層型クラスタリング手法をもちいて、その際に必要となる初期クラスタ重心として、事前に分類された分類結果をもちいることができるので、操作者の意図を反映する文書分類をおこなうことが可能である。

【００４３】また、請求項１７に係る文書分類方法は、文書データを入力する第１工程と、前記第１工程により入力された文書データを解析し解析情報を得る第２工程と、前記第２工程により得られた解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成する第３工程と、前記第３工程により生成された文書特徴ベクトルが文書特徴ベクトル間の類似性を反映する空間に射影されるための表現空間変換関数を算出する第４工程と、前記第４工程により算出された表現空間変換関数をもちいて前記第３工程により生成された文書特徴ベクトルを変換する第５工程と、前記第５工程により変換された文書特徴ベクトル間の類似度に基づいて文書を分類する第６工程と、前記第６工程分類手段により分類された文書分類の結果を記憶する第７工程と、を含んだことを特徴とする。

【００４４】この請求項１７の発明によれば、分類対象である文書群での文書間の類似性に基づいて、各文書をそれら文書間の意味的な関連性を反映しうる表現空間へ変換するための表現空間変換関数を算出し、その表現空間で文書分類をおこなうことにより、操作者の意図を反映しうる文書分類を実現することが可能である。

【００４５】また、請求項１８に係る文書分類方法は、請求項１７の発明において、前記第３工程により生成された文書特徴ベクトル間の内積を算出する第８工程を含み、前記第４工程は、前記第８工程により算出された内積をもちいて表現空間変換関数を算出することを特徴とする。

【００４６】この請求項１８の発明によれば、表現空間変換関数を導出する際に必要となる文書間の類似性として文書特徴ベクトル間の内積をもちいることにより、文書間の意味的な関連性を反映した文書分類をおこなうことが可能である。

【００４７】また、請求項１９に係る文書分類方法は、請求項１８の発明において、前記第１工程により入力された文書の作成者、作成日等の文書データの文書間類似情報を設定する第９工程を含み、前記第４工程は、前記第８工程により算出された内積および前記第９工程により設定された文書間類似情報をもちいて表現空間変換関数を算出することを特徴とする。

【００４８】この請求項１９の発明によれば、表現空間変換関数を導出する際に必要となる文書間の類似性として文書特徴ベクトル間の内積に加え、文書の作成者や作成日などの文書間類似情報ももちいることにより、文書間の意味的な関連性を反映した文書分類をおこなうことが可能である。

【００４９】また、請求項２０に係る文書分類方法は、請求項１７～１９のいずれか一つの発明において、さらに、前記第３工程により生成された文書特徴ベクトルを記憶する第１０工程と、前記第４工程により算出された表現空間変換関数を記憶する第１１工程と、を含んだことを特徴とする。

【００５０】この請求項２０の発明によれば、算出する文書特徴ベクトルと表現空間変換関数を記憶することにより、表現空間変換関数を算出する部分と実際に前記表現空間変換関数をもちいて変換された文書をもちいて文書分類をおこなう部分とを分離して処理するので、その都度、表現空間変換関数を算出することなしに文書分類を実行でき、さらに、前記文書特徴ベクトル変換部でもちいる表現空間変換関数として、事前に他の文書特徴ベクトルに基づいて生成された表現空間変換関数をもちいることもできるため、文書分類の繰り返し実行を短時間で効率良くおこなうことが可能である。

【００５１】また、請求項２１に係る文書分類方法は、請求項１７～２０のいずれか一つの発明において、さらに、前記第５工程により文書特徴ベクトルを変更する前に、前記第２工程により抽出される単語が有する特性により構成される規則をもちいて前記文書特徴ベクトルおよび／または文書特徴ベクトルを構成する特徴次元を操作することにより前記文書特徴ベクトルを修正する第１２工程を含んだことを特徴とする。

【００５２】この請求項２１の発明によれば、文書分類の繰り返し実行をおこなう際、個々の分類実行ごとに、文書特徴ベクトルやそれらを構成する特徴次元を操作することで、各分類ごとに異なる単語を削除して文書分類を実行する等の分類対象文書の範囲の変更や分類をおこなう空間の変更をおこなうことが可能である。

【００５３】また、請求項２２に係る文書分類方法は、請求項２１の発明において、前記第１２工程において文書特徴ベクトルを修正することにより特徴次元が変更された場合に、前記変更された特徴次元により第５工程において前記文書特徴ベクトルが適切に変換できるように、前記第４工程により算出された表現空間変換関数を修正する第１３工程を含んだことを特徴とする。

【００５４】この請求項２２の発明によれば、表現空間変換関数が文書特徴ベクトルの内積をに基づいて算出される場合、表現空間変換関数をもちいて変換された文書をもちいて文書分類をおこなう部分において、文書特徴ベクトルやその特徴次元が操作された場合に生じる表現空間変換関数の不整合を簡便に修正することができるので、より適正な文書特徴ベクトルの変換をおこなうことが可能となる。

【００５５】また、請求項２３に係る文書分類方法は、請求項１７～２１のいずれか一つの発明において、さらに、前記表現空間変換関数の特徴次元の操作に関する指

示をする第１４工程と、前記第１４工程により指示された特徴次元の操作に関する指示内容に基づいて、前記表現空間変換関数を修正する第１５工程と、を含んだことを特徴とする。

【００５６】この請求項２３の発明によれば、前記表現空間変換関数をもちいて構成される空間の特徴次元について操作者が簡便な操作をすることにより、操作者の意図を反映しうる文書分類をおこなうことが可能である。

【００５７】また、請求項２４に係る文書分類方法は、請求項２３の発明において、前記第１５工程により指示された特徴次元の操作に関する指示内容が、任意の文書ベクトルデータをもちいて前記表現空間変換関数の特徴次元を操作するものであることを特徴とする。

【００５８】この請求項２４の発明によれば、前記表現空間変換関数をもちいて構成される空間の特徴次元について、操作者により指示された分類対象以外の任意の文書ベクトルデータをもちいての簡便な操作をすることにより、操作者の意図を反映しうる文書分類をおこなうことが可能である。

【００５９】また、請求項２５に係る文書分類方法は、請求項２３の発明において、前記第１５工程により指示された特徴次元の操作に関する指示内容が、文書特徴ベクトルをもちいて前記表現空間変換関数の特徴次元を操作するものであることを特徴とする。

【００６０】この請求項２５の発明によれば、前記表現空間変換関数をもちいて構成される空間の特徴次元について、操作者により指示された文書特徴ベクトルをもちいての簡便な操作をすることにより、操作者の意図を反映しうる文書分類をおこなうことが可能である。

【００６１】また、請求項２６に係る文書分類方法は、請求項２３の発明において、前記第１５工程により指示された特徴次元の操作に関する指示内容が、前記第２工程により得られた解析情報をもちいて前記表現空間変換関数の特徴次元を操作するものであることを特徴とする。

【００６２】この請求項２６の発明によれば、前記表現空間変換関数をもちいて構成される空間の特徴次元を、操作者により指示された解析情報をもちいての簡便な操作をすることにより、操作者の意図を反映しうる文書分類をおこなうことが可能である。

【００６３】また、請求項２７に係る文書分類方法は、請求項２３の発明において、前記第１５工程により指示された特徴次元の操作に関する指示内容が、前記第７工程により記憶された分類結果をもちいて前記表現空間変換関数の特徴次元を操作するものであることを特徴とする。

【００６４】この請求項２７の発明によれば、前記表現空間変換関数をもちいて構成される空間の特徴次元を、操作者により指示された事前に分類された分類結果をもちいての簡便な操作をすることにより、操作者の意図を反映しうる文書分類をおこなうことが可能である。

【００６５】また、請求項２８に係る文書分類方法は、請求項１７～２７のいずれか一つの発明において、初期クラスタ重心を指定する第１６工程と、前記第１６工程により指定された初期クラスタ重心を登録する第１７工程とを含み、前記第６工程は、前記第１７工程により登録された初期クラスタ重心にしたがって文書を分類することを特徴とする。

【００６６】この請求項２８の発明によれば、文書分類手法として、非階層型クラスタリング手法をもちいて、その際に必要となる初期クラスタ重心を、操作者が任意に指定することができ、その指定された初期クラスタ重心にしたがって文書分類をおこなうので、操作者の意図を反映する文書分類をおこなうことが可能である。

【００６７】また、請求項２９に係る文書分類方法は、請求項２８の発明において、前記第１６工程により指定される初期クラスタ重心として任意の文書ベクトルデータを指定することを特徴とする。

【００６８】この請求項２９の発明によれば、文書分類手法として、非階層型クラスタリング手法をもちいて、その際に必要となる初期クラスタ重心として、分類対象以外の任意の文書をもちいることができるので、操作者の意図を反映する文書分類をおこなうことが可能である。

【００６９】また、請求項３０に係る文書分類方法は、請求項２８の発明において、前記第１６工程により指定される初期クラスタ重心として文書特徴ベクトルを指定することを特徴とする。

【００７０】この請求項３０の発明によれば、文書分類手法として、非階層型クラスタリング手法をもちいて、その際に必要となる初期クラスタ重心として、文書特徴ベクトルをもちいることができるので、操作者の意図を反映する文書分類をおこなうことが可能である。

【００７１】また、請求項３１に係る文書分類方法は、請求項２８の発明において、前記第１６工程により指定される初期クラスタ重心として前記第２工程により得られた解析情報を指定することを特徴とする。

【００７２】この請求項３１の発明によれば、文書分類手法として、非階層型クラスタリング手法をもちいて、その際に必要となる初期クラスタ重心として、分類対象文書を文書解析部に作用させた結果得られる単語等の解析情報をもちいることができるので、操作者の意図を反映する文書分類をおこなうことが可能である。

【００７３】また、請求項３２に係る文書分類方法は、請求項２８の発明において、前記第１６工程により指定される初期クラスタ重心として前記第７工程により記憶された分類結果を指定することを特徴とする。

【００７４】この請求項３２の発明によれば、文書分類手法として、非階層型クラスタリング手法をもちいて、その際に必要となる初期クラスタ重心として、事前に分

類された分類結果をもちいることができるので、操作者の意図を反映する文書分類をおこなうことが可能である。

【００７５】また、請求項３３の発明に係る記憶媒体は、請求項１７～３２に記載された方法をコンピュータに実行させるプログラムを記録したことで、そのプログラムを機械読み取り可能となり、これによって、請求項１７～３２の動作をコンピュータによって実現することが可能である。

【００７６】

【発明の実施の形態】以下に添付図面を参照して、この発明に係る文書分類装置、文書分類方法およびその方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体の好適な実施の形態を詳細に説明する。

【００７７】（実施の形態１）まず、この発明の実施の形態１による文書分類装置を構成する情報処理システム全体のハードウエア構成を説明する。図１は、実施の形態１による文書分類装置を構成する情報処理システム全体のハードウエア構成を示す説明図である。

【００７８】図１において、実施の形態１による文書分類装置を構成する情報処理システムは、サーバー／クライアント方式で構成されている。すなわち、サーバー１０１と複数のクライアント１０２がネットワーク１０３によって接続されている。クライアント１０２は、分類データの生成、サーバー１０１への指示、分類結果の表示などをおこなう。一方、クライアント１０２からの指示にしたがって、サーバー１０１は文書（テキスト）分類に関する処理を膨大な数値演算によりおこない、その処理の結果をクライアント１０２へ送る。

【００７９】より具体的には、サーバー１０１においては、テキスト分類処理（前処理、クラスタリング処理）がおこなわれ、クライアント１０２においては、分類データ生成、処理実行指示、テキスト分類結果表示等がおこなわれる。サーバー１０１における処理は、上述のように、「前処理」と「分類処理」の２つに分かれており、その処理はデータによっては非常に負荷が大きくなる。したがって、サーバー１０１は「前処理」と「分類処理」がそれぞれ一つずつしか処理をおこなわないようにマネージャプロセスが処理受付リストを作成して管理する。「前処理」および「分類処理」の詳細については後述する。

【００８０】また、サーバー１０１とクライアント１０２との間のデータのやりとりはファイル共有という方法をもちいる。すなわち、分類処理にもちいるファイルをサーバー１０１上の共有フォルダに作成することにより両者はデータのやりとりをおこなう。したがって、クライアント１０２からはサーバー１０１の共有フォルダをネットワーク共有して利用することが可能である。

【００８１】つぎに、サーバー１０１およびクライアント１０２のハードウエア構成について説明する。図２は、実施の形態１による文書分類装置を構成する情報処理システムにおけるサーバー１０１をハードウエア的に示す説明図である。サーバー１０１は、たとえばワークステーション（ＷＳ）等がもちいられる。

【００８２】図２において、２０１はサーバー１０１全体を制御するＣＰＵを、２０２はブートプログラム等を記憶したＲＯＭを、２０３はＣＰＵ２０１のワークエリアとして使用されるＲＡＭ２０３を、２０４は通信回線２０５を介してネットワーク１０３に接続され、そのネットワーク１０３と内部のインターフェイスを司るインターフェイス（Ｉ／Ｆ）を、２０６はデータを記憶するディスク装置を示している。２００は上記各部を結合させるためのバスを示している。

【００８３】そのほか、文書情報、画像情報、機能情報等を表示するディスプレイ２０８や，データを入力するためのキーボード２０９およびマウス２１０等が同様に接続されていてもよい。さらに、ディスク装置２０６には、クライアント１０２との間のデータのやりとりをするための共有フォルダ２０７が設けられている。

【００８４】また、図３は、実施の形態１による文書分類装置を構成する情報処理システムにおけるクライアント１０２をハードウエア的に示す説明図である。クライアント１０２は、たとえばパーソナルコンピュータ（ＰＣ）等がもちいられる。

【００８５】図３において、３０１はシステム全体を制御するＣＰＵを、３０２はブートプログラム等を記憶したＲＯＭを、３０３はＣＰＵ３０１のワークエリアとして使用されるＲＡＭを、３０４はＣＰＵ３０１の制御にしたがってＨＤ（ハードディスク）３０５に対するデータのリード／ライトを制御するＨＤＤ（ハードディスクドライブ）を、３０５はＨＤＤ３０４の制御で書き込まれたデータを記憶するＨＤを、３０６はＣＰＵ３０１の制御にしたがってＦＤ（フロッピーディスク）３０７に対するデータのリード／ライトを制御するＦＤＤ（フロッピーディスクドライブ）を、３０７はＦＤＤ３０６の制御で書き込まれたデータを記憶する着脱自在のＦＤを、３０８はドキュメント、画像、機能情報等を表示するディスプレイをそれぞれ示している。

【００８６】また、３０９は通信回線３１０を介してネットワーク１０３に接続され、そのネットワーク１０３と内部のインターフェイスを司るインターフェイス（Ｉ／Ｆ）を、３１１は文字、数値、各種指示等の入力のためのキーを備えたキーボードを、３１２はカーソルの移動や範囲選択、あるいは表示画面に表示されたアイコンやボタンの押下やウインドウの移動やサイズの変更等をおこなうマウスを、３１３はＯＣＲ（Ｏｐｔｉｃａｌ Ｃｈａｒａｃｔｅｒ Ｒｅａｄｅｒ）機能を備えた画像を光学的に読み取るスキャナを、３１４は分類結果を含むデータの内容等を印刷するプリンタを、３１５は上記

各部を結合するためのバスをそれぞれ示している。また、HD３０５にはワープロソフトや表計算ソフト等のアプリケーションソフト３１６が記憶されている。

【００８７】つぎに、実施の形態１による文書分類装置の機能的構成について説明する。図４～図６は、実施の形態１による文書分類装置の構成を機能的に示すブロック図である。図４において、文書分類装置は、入力部４０１と、解析部４０２と、ベクトル生成部４０３と、変換関数算出部４０４と、ベクトル変換部４０５と、分類部４０６と、分類結果記憶部４０７を含む構成である。

【００８８】さらに、入力部４０１と解析部４０２との間には、文書データ中の表記の揺れ等を吸収する図示しない第１フィルタ部を含めるようにしてもよい。また、解析部４０２とベクトル生成部４０３との間には、解析情報から不要な単語や品詞を除去する図示しない第２フィルタ部を含めるようにしてもよい。さらに、変換関数算出部４０４とベクトル変換部との間には、文書特徴ベクトルから分類時に不要な単語や品詞を除去する図示しない第３フィルタ部を含めるようにしてもよい。

【００８９】また、図５においては、さらに内積算出部４２１を含む構成となっている。また、図６においては、さらに文書間類似情報設定部４３１を含む構成となっている。

【００９０】入力部４０１は、文書データを入力するものであり、たとえば、キーボード２０９または３１１、スキャナ３１３、OCR機能を備えたスキャナ３１３、またはネットワーク１０３を経由して文書や文書群を得ることができるＩ／Ｆ２０４または３０９等である。また、入力部４０１は、上記以外に、文書データを取得することができるものであれば、それらのすべてを含む。たとえば、文書データがデータベース化されている場合に、そのデータベースが記録された媒体を本実施の形態の文書分類装置に組み入れた場合も文書データの入力とする。さらに、入力した文書データを記憶する図示しない文書データ記憶部を含んでいてもよい。

【００９１】ここで、文書とは、自然言語で記述された一つ以上の文の集まりであり、それが分類対象となる場合はこれを文書という。具体的には、公開特許公報や特定の新聞記事も文書であり、また、請求項や特定の一文を取り出したものであっても、これを文書とみなすものである。

【００９２】解析部４０２は、入力部４０１により入力された文書データの単語を解析し解析情報を得る。具体的には、入力部４０１により入力された文書データそれぞれに対して、形態素解析等の自然言語解析をおこない、単語やその品詞などを抽出する。さらに、文書群で出現した単語に対し一意な単語ＩＤを付与し、文書内および文書群に対する単語出現回数を計数するものである。

【００９３】また、ベクトル生成部４０３は、解析部４０３により得られた解析情報に基づいて文書データに対する文書特徴ベクトルを生成するものである。変換関数算出部４０４は、文書特徴ベクトル間の類似性を反映する空間に前記ベクトル生成部４０３により生成された文書特徴ベクトルを射影するための表現空間変換関数を算出するものである。ベクトル変換部４０５は、変換関数算出部４０４により算出された表現空間変換関数をもちいて文書特徴ベクトルを変換するものである。

【００９４】ベクトル生成部４０３，変換関数算出部４０４，ベクトル変換部４０５の各処理の詳細は後述する。

【００９５】分類部４０６は、ベクトル変換部４０５により変換された新たな文書特徴ベクトル間の類似度に基づいて文書を分類するものである。具体的には、生成された分類対象データに対して、カイ自乗法の手法、判別分析の手法、およびクラスタ分析の手法等の分類手法を適用することで、文書分類をおこなうことができる。分類部４０６においては、ベクトルデータが適用できる分類手法であれば、その手法は問わない。

【００９６】さらに、分類結果記憶部４０７は、分類部４０６により分類された結果を適切な形式で記憶する記憶部である。たとえば、ディスク装置３０６またはハードディスク３１６の所定の領域のほか、RAM２０３または３０３、その他データを記憶可能なところであればいずれでもよい。

【００９７】内積算出部５０１は、ベクトル生成部４０２手段により生成された文書特徴ベクトル間の内積を算出する算出部である。内積算出部５０１の処理の内容は後述する。

【００９８】また、文書間類似情報設定部６０１は、入力部４０１により入力された文書の作成者、作成日等の文書データの文書間類似情報を設定する設定部である。文書間類似情報には、文書内での単語の出現順序や、文書の作成日、修正日、作成者、修正者、参照文書、引用文書などの文書間での一致情報を含む。操作者は、こらの文書間類似情報の中から所望の情報を指定し任意に設定することができる。

【００９９】入力部４０１、解析部４０２、ベクトル生成部４０３、変換関数算出部４０４、ベクトル変換部４０５、分類部４０６、内積算出部４２０、文書間類似情報設定部４３０は、ROM２０２または３０２、RAM２０３または３０３、あるいはディスク装置３０６またはハードディスク３１６等の記録媒体に記録されたプログラムに記載された命令にしたがってＣＰＵ２０１または３０１等が命令処理を実行することにより、各部の機能を実現する。

【０１００】つぎに、ベクトル生成部４０３による文書特徴ベクトルの生成処理の内容について説明する。ベクトル生成部４０３は、解析部４０３により得られた解析情報に基づいて文書データに対する文書特徴ベクトルを

生成するものである。ここで解析情報とは、たとえば、単語、単語ＩＤ、単語出現回数、品詞情報等の情報である。

【０１０１】図７は、文書－単語行列データと文書特徴ベクトルの一例を示す説明図である。図７において、行成分７０１が単語ＩＤであり、また、列成分７０２が文書ＩＤである。行列要素として、文書ＩＤが列番号であり、文書に含まれる単語ＩＤが行番号である単語の出現回数となるような文書－単語行列データを上記解析情報に基づいて生成する。この文書－単語行列の各列ベクトルが文書特徴ベクトルである。このようにして文書特徴ベクトルを生成する。

【０１０２】また、この文書特徴ベクトルに対して、正規化等の処理を同時におこなうこともできる。この際、文書－単語行列データに付随する付加的な情報、たとえば、文書－単語行列データの行成分である単語ＩＤとその単語との対応関係を記述した単語－単語ＩＤ対応マップデータや各単語において単語ＩＤとその単語が有する品詞情報との対応関係を記述した単語ＩＤ－品詞対応マップデータなども同時に生成する。

【０１０３】つぎに、変換関数算出部４０４による変換関数算出処理の内容について説明する。ベクトル生成部４０３における文書特徴ベクトルの生成は、通常、その文書内での単語の出現回数に基づいておこなわれる。この際、各単語はそれぞれ意味的に独立なものと仮定し、各々を直交するものとして扱われる。しかしながら、現実には単語は多義性や同義性を含むものであるため、上記のような仮定の妥当性は保証されておらず、各単語が各々直交するものと扱われることにより、分類の精度・妥当性にも影響を及ぼすものである。

【０１０４】この影響を軽減するための手法として、この問題を多次元尺度問題とみなして、統計的手法をもちいることが考えられる。すなわち、変換関数算出部４０４において、各文書特徴ベクトルを文書特徴ベクトル間での特徴次元、すなわち単語の共起性が反映された空間へ変換するための表現空間変換関数を、ベクトル生成部４０３により生成された文書特徴ベクトルに基づいて算出する。なお、単語間の同義性の影響を軽減するための方法としてシソーラス等をもちいるようにしてもよい。

【０１０５】本実施の形態においては表現空間変換関数の算出手法としては、前出の「Ｒｅｐｒｅｓｅｎｔａｔｉｎｇ　Ｄｏｃｕｍｅｎｔｓ　Ｕｓｉｎｇ　ａｎ　Ｅｘｐｌｉｃｉｔ　Ｍｏｄｅｌ　ｏｆ　Ｔｈｅｉｒ　Ｓｉｍｉｌａｒｉｔｉｅｓ」に述べられている表現空間変換関数の算出手法をもちいるが、そのほか、因子分析や数量化などの手法をもちいて算出するようにしてもよい。

【０１０６】すなわち、内積算出部５０１により算出された文書特徴ベクトル間の内積に、文書間類似情報設定手段により設定された文書間類似情報を付加した文書間類似行列と文書特徴ベクトルに基づいて表現空間変換関数を算出する。そして、この表現空間変換関数をもちいることにより、文書間の意味的な類似性を強く反映した表現空間にて文書分類をおこなうことができる。また、上述のように、操作者が自由に文書間類似情報を付加的に選択することもできるため、操作者の意図を反映した文書分類をおこなうことができる。

【０１０７】具体的には、文書数をｄ、単語数をｔとし、大きさｔ×ｄの文書－単語行列（文書特徴ベクトル）をＸ、大きさｄ×ｄの文書間内積行列をＳ、大きさｄ×ｄの付加的文書間類似情報行列をS_aとすると、表現空間変換関数Ｗは式１のようになる。

【０１０８】$W=M^T C X^+$　（式１）

【０１０９】なお、^{T}は行列の転置を示す。

【０１１０】ここで、行列へ特異値分解を適用する演算子をｓｖｄ（　）とすると、式１において、行列Ｃ、Ｍ、X^+はつぎのような行列となる。

【０１１１】$X=svd(X)=ULA^T$　（式２）

【０１１２】$S=X^T X$　（式３）

【０１１３】

$S+S_a=svd(S+S_a)=C^T C$　（式４）

【０１１４】

$CAA^T=svd(CAA^T)=MZN^T$　（式５）

【０１１５】$X^+=AL^{-1}U^T$　（式６）

【０１１６】また、ベクトルの内積をもちいて表現空間変換関数を算出するには、上記付加的文書間類似行列S_aを空行列とする。その場合、表現空間変換関数は式７にようになる。

【０１１７】$W=U^T$　（式７）

【０１１８】また、設定された文書間類似情報をもちいて表現空間変換関数を算出するには、S_aを空行列以外の対称行列とする。

【０１１９】さらに、本文書分類装置では、表現空間変換関数Ｗを大きさがｔ×ｔの単位行列とすることで変換関数生成部４０４をバイパスすることも可能である。

【０１２０】さらにまた、ベクトル生成部４０３で生成される文書特徴ベクトルは、特徴次元数が文書群で出現する単語数であるため、通常非常に高次なものとなり、このまま分類等をおこなうと計算コストや記憶空間が膨大になる。このため、出現回数の極端に少ない単語や極端に多い単語を文書特徴ベクトルを構成する次元から除外することができるが、これにより分類精度や妥当性が低下する可能性がある。

【０１２１】本発明でもちいる表現空間変換関数は各文書特徴ベクトル間の単語の共起性が考慮された空間への変換を実現するため、式１からも明らかなように表現空間変換関数により生成される表現空間は各特徴次元が複数の単語の一次結合として表現される。したがって、少ない特徴次元でも多くの単語の意味を扱うことができ、これにより分類等をおこなう際の計算コストや記憶空間を抑制することができる。

【0122】つぎに、ベクトル変換部405による文書特徴ベクトルの変換処理について説明する。ベクトル変換部405では、変換関数生成部404で生成される表現空間変換関数をもちいて、文書特徴ベクトルを変換し、分類の対象となるデータを導く。加えて、各単語も前記表現空間変換関数をもちいて変換するが可能である。すなわち、表現空間変換関数として行列Wをもちいると、変換された文書特徴ベクトルをD_pとすると、式8のようになる。

【0123】$D_p = WX$　（式8）

【0124】また、変換された単語の行列表現をT_pとすると、式9のようになる。

【0125】$T_p = W^T I = W$　（式9）

【0126】なお、Ｉは単位行列を示す。

【0127】つぎに、実施の形態１による文書分類装置の一連の処理の手順について説明する。図８は実施の形態１による文書分類装置の一連の処理の手順を示すフローチャートである。図８のフローチャートにおいて、まず、入力部４０１は文書データを入力する（ステップＳ８１０）。つぎに、解析部４０２はステップＳ８１０において入力された文書データを解析し解析情報を得る（ステップＳ８２０）。

【0128】つぎに、ベクトル生成部４０３はステップＳ８２０において得解析情報に基づいて文書特徴ベクトルを生成する（ステップＳ８３０）。つぎに、変換関数算出部４０４はステップ８３０において生成された文書特徴ベクトルが文書特徴ベクトル間の類似性を反映する空間に射影されるための表現空間関数を算出する（ステップＳ８４０）。

【0129】つぎに、ベクトル変換部４０５はステップＳ８４０において算出された表現空間関数をもちいてステップＳ８３０において生成された文書特徴ベクトルを変換する（ステップＳ８５０）。つぎに、分類部４０６はステップＳ８５０において変換された文書特徴ベクトルの間の類似度に基づいて文書を分類する（ステップＳ８６０）。その後、ステップＳ８６０によって分類された分類結果が記憶され（ステップＳ８７０）、すべての処理を終了する。

【0130】また、図９は実施の形態１による文書分類装置の一連の処理の別の手順を示すフローチャートである。図９のフローチャートにおいて、図８の各ステップと同じ処理をおこなうステップは同じ番号を付して、その説明を省略する。

【0131】ステップＳ８３０につづいて、同ステップにおいて生成された文書特徴ベクトル間の内積を算出する（ステップＳ８３５）。つぎに、文書間類似情報をもちいるとの指示があったか否かを判断する（ステップＳ８３６）。

【0132】ステップＳ８３６において、指示がなかった場合（ステップＳ８３６否定）は、ステップＳ８３５において算出された内積をもちいて表現空間変換関数の算出をする（ステップＳ８４０）。一方、ステップＳ８３６において、指示があった場合（ステップＳ８３６肯定）は、入力部４０１により入力された文書データの文書間類似情報を設定する（ステップＳ８３７）。その後、ステップＳ８４０へ移行し、ステップＳ８３５において算出された内積とステップＳ８３７において設定された文書類間情報をもちいて表現空間変換関数の算出をする。以下、図８と同様の処理をおこなう。

【0133】以上説明したように、実施の形態１によれば、分類対象である文書群での文書間の類似性に基づいて、各文書をそれら文書間の意味的な関連性を反映しうる表現空間へ変換するための表現空間変換関数を算出し、その表現空間で文書分類をおこなうことにより、操作者の意図を反映しうる文書分類を実現することができる。

【0134】（実施の形態２）さて、上述した実施の形態１では、生成された文書特徴ベクトルと算出された表現空間変換関数の保存についてはなんら記載していなかったが、以下に説明する実施の形態２のように、さらにベクトル記憶部と、変換関数記憶部とを含む構成とするようにしてもよい。

【0135】実施の形態２による文書分類装置の機能的構成について説明する。図１０は、実施の形態２による文書分類装置の構成を機能的に示すブロック図である。図１０において、実施の形態１の図４と同一のものに関しては同じ番号を付して、その説明を省略する。

【0136】ベクトル記憶部１００１は、ベクトル生成部４０３により生成された文書特徴ベクトルを記憶する記憶部である。この際、ベクトル生成部４０３において同時に生成される文書－単語行列データに付随する付加的な情報、たとえば、文書－単語行列データの行成分である単語ＩＤとその単語との対応関係を記述した単語－単語ＩＤ対応マップデータや各単語において単語ＩＤとその単語が有する品詞情報との対応関係を記述した単語ＩＤ－品詞対応マップデータや構文情報データなども記憶することができる。

【0137】また、変換関数記憶部１００２は、変換関数生成部４０４より生成された表現空間変換関数を記憶する記憶部である。

【0138】ベクトル記憶部１００１、変換関数記憶部１００２は、ＲＯＭ２０２または３０２、ＲＡＭ２０３または３０３、あるいはディスク装置３０６またはハードディスク３１６等の記録媒体に記録されたプログラムに記載された命令にしたがってＣＰＵ２０１または３０１等が命令処理を実行することにより、各部の機能を実現する。

【0139】文書特徴ベクトルと表現空間変換関数を記憶することにより記憶された文書表現空間をもちいて記憶された文書特徴ベクトルを変換することが可能になる

ため、ベクトル記憶部１００１および変換関数記憶部１００２と、ベクトル変換部４０５を一連の処理としておこなう必要がなくなり、機能的に分離することができる。

【０１４０】つぎに、実施の形態２による文書分類装置の一連の処理の手順について説明する。図１１は実施の形態２による文書分類装置の一連の処理の手順を示すフローチャートである。図１１のフローチャートにおいて、実施の形態１の図８の各ステップと同じ処理をおこなうステップは同じ番号を付して、その説明を省略する。

【０１４１】ステップＳ８３０の処理につづいて、同ステップにおいて生成された文書特徴ベクトルを記憶する（ステップＳ８３１）。その後、ステップＳ８４０へ移行し、実施の形態１と同様の処理をおこなう。また、ステップＳ８４０の処理につづいて、同ステップにおいて算出された表現空間変換関数を記憶する（ステップＳ８４１）。以下、実施の形態１と同様の処理をおこなう。

【０１４２】以上説明したように、実施の形態２による文書分類装置は、分類数や分類手法を変えて分類をおこなう場合に、その都度、表現空間変換関数を算出することなしに文書分類を実行できるため、複数の分類結果を短時間で得ることができる。

【０１４３】さらに、前記文書特徴ベクトル変換部でもちいる表現空間変換関数として、事前に他の文書特徴ベクトルに基づいて生成された表現空間変換関数をもちいることもできる。

【０１４４】（実施の形態３）さて、実施の形態１，２に対して、以下に説明する実施の形態３のように、さらにベクトル修正部１２０１を追加してもよい。

【０１４５】まず、実施の形態３による文書分類装置の機能的構成について説明する。図１２は、実施の形態３による文書分類装置の構成を機能的に示すブロック図である。図１２において、実施の形態１の図４と同一のものに関しては同じ番号を付して、その説明を省略する。

【０１４６】ベクトル修正部１２０１は、ベクトル変換部４０５により文書特徴ベクトルを変更する前に、解析部４０２により抽出される単語が有する特性により構成される規則をもちいて文書特徴ベクトル、文書特徴ベクトルを構成する特徴次元のいずれか一つまたはその両方を操作することにより文書特徴ベクトルを修正するものである。

【０１４７】図１３はベクトル修正部１２０１の処理内容の手順を示すフローチャートである。図１３のフローチャートにおいて、ベクトル修正部１２０１は、まず、文書特徴ベクトルの読み込みをおこない（ステップＳ１３０１）、つぎに、解析部４０２おいて抽出された単語やその単語の品詞情報などを指定することで（ステップＳ１３０２）、削除などの操作をおこなう前記文書特徴ベクトルの特徴次元、すなわち文書群に固有に出現する単語の単語ＩＤを決定する（ステップＳ１３０３）。

【０１４８】その後、ベクトル生成部４０３よって生成された文書特徴ベクトルやベクトル記憶部１００１によって記憶された文書特徴ベクトルに対して、操作対象の特徴次元に対し、削除や合成等の修正の操作をおこない（ステップＳ１３０４）、文書特徴ベクトルを合成する。

【０１４９】ベクトル修正部１２０１は、ＲＯＭ２０２または３０２、ＲＡＭ２０３または３０３、あるいはディスク装置３０６またはハードディスク３１６等の記録媒体に記録されたプログラムに記載された命令にしたがってＣＰＵ２０１または３０１等が命令処理を実行することにより、各部の機能を実現する。

【０１５０】文書特徴ベクトルからｔ’個の特徴次元（すなわち、単語ＩＤ）を削除する手続きの一例を図１４に示す。文書数ｄ、単語数ｔとし、文書特徴ベクトル（文書－単語頻度行列）をｔ×ｄの大きさの行列Ｘとし、各行（列）が単語ＩＤに対応する大きさｔ×ｔの単位行列に、削除対象となる単語ＩＤに対応する行を削除したｔ’×ｔの大きさの行列をP_tとした場合、修正部１２０１によって修正される文書特徴ベクトルＸ’は式１０のようになる。

【０１５１】$X' = P_t X$　（式１０）

【０１５２】つぎに、実施の形態３による文書分類装置の一連の処理の手順について説明する。図１５は実施の形態１による文書分類装置の一連の処理の手順を示すフローチャートである。図１３のフローチャートにおいて、実施の形態１の図８の各ステップと同じ処理をおこなうステップは同じ番号を付して、その説明を省略する。

【０１５３】ステップＳ８３０、Ｓ８３１の処理につづいて、文書特徴ベクトルの修正をおこなう（ステップＳ８３２）。その後、ステップＳ８４０へ移行し、実施の形態１と同様の処理をおこなう。

【０１５４】以上説明したように、実施の形態３による文書分類装置は、ベクトル修正部１２０１により、ベクトル生成部４０３によって文書特徴ベクトルを生成した後でも、文類時に不要であることが判明した単語などを削除することができる。さらに、同じ文書特徴ベクトルに対しての分類を効率的におこなるようになっているが、前記文書特徴ベクトル修正部１２０１により、各分類ごとに異なる単語を削除して文書分類を実行することができる。

【０１５５】（実施の形態４）さて、実施の形態３では、ベクトル修正部１２０１を追加したが、以下に説明する実施の形態４のように、さらに、ベクトル修正部とともに変換関数修正部１６０１も併せて追加してもよい。

【０１５６】まず、実施の形態４による文書分類装置の機能的構成について説明する。図１６は、実施の形態４

による文書分類装置の構成を機能的に示すブロック図である。図１６において、実施の形態３の図１２と同一のものに関しては同じ番号を付して、その説明を省略する。

【０１５７】実施の形態３において、ベクトル修正部１２０１によって文書特徴ベクトルの修正がおこなわれた場合に、表現空間辺関数は修正前の文書特徴ベクトルに基づいて算出されているため、この表現空間変換関数にも文書特徴ベクトルが修正された効果を反映させなければ、文書特徴ベクトルを修正した効果が半減する可能性がある。したがって、前記表現空間変換関数を修正された文書特徴ベクトルをもとに修正する。

【０１５８】すなわち、図１６における変換関数修正部１６０１は表現空間変換関数をW'に修正する。表現空間変換関数が文書特徴ベクトルの内積に基づいて算出される場合に、表現空間変換関数は式７で与えられる。このとき、修正された表現空間変換関数をW'とすると、式2、式7、式１０をもちいて式１１のように表現される。

【０１５９】

$$W' = L^{-1}U^{T} P_t X (P_t X) \quad (式11)$$

【０１６０】変換関数修正部１６０１は、ＲＯＭ２０２または３０２、ＲＡＭ２０３または３０３、あるいはディスク装置３０６またはハードディスク３１６等の記録媒体に記録されたプログラムに記載された命令にしたがってＣＰＵ２０１または３０１等が命令処理を実行することにより、各部の機能を実現する。

【０１６１】図１７に実施の形態４による文書分類装置の一連の処理の手順を説明するフローチャートを示す。図１７のフローチャートにおいて、文書特徴ベクトルの修正があった場合、表現空間変換関数の修正もおこなう（ステップＳ８４１）。以下は実施の形態３の処理と同様である。

【０１６２】以上説明したように、実施の形態４による文書分類装置においては、文書特徴ベクトルの修正にともなって表現空間変換関数の修正もおこなうことができるので、より適正な文書特徴ベクトルの変換ができる。

【０１６３】（実施の形態５）さて、実施の形態４では、変換関数修正部１６０１を追加したが、以下に説明する実施の形態５のように、さらに変換関数修正部１６０１へ修正指示をおこなう変換関数修正指示部１８０１を追加してもよい。

【０１６４】まず、実施の形態５による文書分類装置の機能的構成について説明する。図１８は、実施の形態５による文書分類装置の構成を機能的に示すブロック図である。図１８において、実施の形態１の図４と同一のものに関しては同じ番号を付して、その説明を省略する。

【０１６５】変換関数修正指示部１８０１は、表現空間変換関数の特徴次元の操作に関する指示するものである。また、変換関数修正部１８０２は、変換関数修正指示部１８０１からの指示内容に基づいて、前記表現空間変換関数の特徴次元を操作し、前記表現空間変換関数を修正する。

【０１６６】変換関数修正指示部１８０１、変換関数修正部１８０２は、ＲＯＭ２０２または３０２、ＲＡＭ２０３または３０３、あるいはディスク装置３０６またはハードディスク３１６等の記録媒体に記録されたプログラムに記載された命令にしたがってＣＰＵ２０１または３０１等が命令処理を実行することにより、各部の機能を実現する。

【０１６７】変換関数修正指示部１８０１においては、操作者の意図を反映するような文書分類をおこなうための一つの方法として、前記表現空間変換関数により構成される空間における不必要な特徴次元や、悪影響を及ぼすような特徴次元に対し削除や合成をおこなったり、逆にある特徴次元を強調させるための操作をすることが考えられる。

【０１６８】しかしながら、表現空間変換関数により生成される空間の特徴次元は、解析部４０２によって抽出された単語のうち意味的に似たものが複数結合したものと考えることができるため、各特徴次元の意味的な解釈は極めて複雑かつ多義的なものである。したがって、操作者に各特徴次元の意味を提示することは極めて難しい。

【０１６９】そこで、操作者に分類に反映させたくない内容や強調したい内容をもつ文書や単語などの情報を指定させ、それらを前記表現空間辺関数により構成される空間に適切に射影し、それらと類似度の高い特徴次元や低い特徴次元を判別することで、操作をおこなう特徴次元を選択することができる。

【０１７０】本実施の形態では、表現空間変換関数の特徴次元を操作する例として、操作者が指定するある文書と類似度の高い特徴次元の削除をおこなう例を示す。操作者により指定された文書を前記文書特徴ベクトルと同じ次元数をもつベクトルで表現し、その文書ベクトルに表現空間変換関数を適用し文書ベクトルを表現空間変換関数により構成される空間へ射影する。そして、この射影された文書ベクトルと各特徴次元との類似度を算出することで、類似度の高い特徴次元を判別する。

【０１７１】このとき、類似度を測るための尺度としては、余弦尺度、内積尺度、ユークリッド距離尺度などをもちいることができる。また、判別に関しては、ある類似度以上を削除対象として採用するような閾値処理による判別や、類似度の高い順にある一定数を削除対象として採用する定数処理もしくは判別分析などももちいることができる。

【０１７２】このようにして、採用された特徴次元を表現空間変換関数から削除することで表現空間変換関数を修正することができる。この際、操作者が指示（指定）する情報としては、前記文書特徴ベクトルと同じ次元数

を有するベクトル形式であれば、どのようなものでも適用可能である。

【０１７３】操作者が指示する情報としては、そのほかに、より操作者にとって理解しやすいものとして、分類対象文書群以外の文書を文書特徴ベクトルと同じ次元をもつベクトルに表現したものをもちいることができる。また、操作者が指示する情報としてそのほかには、文書特徴ベクトルをもちいることができる。

【０１７４】また、操作者が指示する情報としてそのほかには、解析部４０２によって抽出されまたは操作者が手動で入力した単語や単語品詞情報をもちいることができる。また、操作者が指示する情報としてそのほかには、分類結果記憶部４０７によって記憶されている事前におこなわれた分類結果である分類代表値をもちいることができる。

【０１７５】上記の指示情報は、それぞれ単独でもちいるほか、それらを適切に組み合わせたものをもちいるようにしてもよい。

【０１７６】図１９に実施の形態５による文書分類装置の一連の処理の一部の手順を説明するフローチャートを示す。図１９のフローチャートにおいて、まず、変換関数の修正の指示があるのを待って（ステップＳ１９０１肯定）、つぎに、指示の内容、すなわち、操作者が指示（指定）した指示情報をインプットする（ステップＳ１９０２）。複数の指示がある場合は、すべての指示が終了するまで同様のステップを繰り返し、指示が終了した場合（ステップＳ１９０３肯定）に、インプットされた指示情報に基づいて変換関数の修正を実行し（ステップＳ１９０４）、すべての処理を終了する。

【０１７７】以上説明したように、この実施の形態５によれば、表現空間変換関数をもちいて構成される空間の特徴次元について操作者が簡便な操作をすることにより、操作者の意図を反映しうる文書分類をおこなうことができる。

【０１７８】（実施の形態６）さて、実施の形態１～５に対して、以下に説明する実施の形態６のように、初期重心指定部２００１および初期重心登録部２００２をさらに追加するようにしてもよい。

【０１７９】まず、実施の形態６による文書分類装置の機能的構成について説明する。図２０は、実施の形態６による文書分類装置の構成を機能的に示すブロック図である。図２０において、実施の形態１の図４と同一のものに関しては同じ番号を付して、その説明を省略する。

【０１８０】初期重心指定部２００１は、初期クラスタ重心を指定する指定部である。初期重心登録部２００２は、初期重心指定部２００１により指定された初期クラスタ重心を登録するする登録部である。また、分類部４０５は、初期重心登録部２００２により登録された初期クラスタ重心にしたがって文書を分類するものである。

【０１８１】初期重心指定部２００１、初期重心登録部２００２は、ＲＯＭ２０２または３０２、ＲＡＭ２０３または３０３、あるいはディスク装置３０６またはハードディスク３１６等の記録媒体に記録されたプログラムに記載された命令にしたがってＣＰＵ２０１または３０１等が命令処理を実行することにより、各部の機能を実現する。

【０１８２】通常、カイ自乗法の手法、判別分析の手法、およびクラスタ分析の手法等をもちいて文書分類をおこなう場合にもちいられる分類基準が統計的な理論を元にして構成されている。しかしながら、本実施の形態においては、文書分類をおこなった際の最終的な分類の質の評価は、統計的な数値評価ではなく、その分類結果を分析する操作者による主観評価となる。したがって、前記文書分類をおこなうための諸手法において、操作者が介入できうる余地を設けることで分類結果に操作者の意図を反映することができ、結果として分類結果の質的な向上が見込まれる。

【０１８３】つぎに、図２１に実施の形態６による文書分類装置の一連の処理の一部の手順を説明するフローチャートを示す。非階層型のクラスタリング手法は一般的に図１９のフローチャートのような処理の手順となる。図２１のフローチャートにおいて、まず、初期クラスタ重心が指定され（ステップＳ２１０１）、その初期クラスタ重心が登録される（ステップＳ２１０２）。つぎに、初期クラスタ重心を決定し（ステップＳ２１０３）、そのクラスタ重心と各分類対象データとの類似度を計算する（ステップＳ２１０４）。

【０１８４】つぎに、各分類対象データを一番類似度の高いクラスタに割り当てて（ステップＳ２１０５）、各クラスタごとに割り当てられた分類対象データを基にそのクラスタ重心を計算する（ステップＳ２１０６）。

【０１８５】この時点で、反復停止基準を満たすか否かを判断し（ステップＳ２１０７）、反復停止基準を満たさない場合（ステップＳ２１０７否定）は、ステップＳ２１０４へ移行し、以後、ステップＳ２１０４～Ｓ２１０６の各ステップを繰り返し実行する。ステップＳ２１０７において、反復停止基準を満たす場合（ステップＳ２１０７肯定）は、すべての処理終了する。

【０１８６】分類結果はどのような初期クラスタ重心を選択するかに強く依存するといわれている。したがって、分類実行部での分類手法として、ｋ－ｍｅａｎｓ法などの非階層型クラスタリング手法をもちいて、その初期クラスタ重心を操作者が指定することで、操作者の分類手続きへの介入を可能にし、操作者の意図を反映した文書分類が実現できる。

【０１８７】なお、各文書特徴ベクトルとクラスタの重心ベクトルとの類似度を算出し、各特徴ベクトルで最も類似度の高い分類代表値にその文書特徴ベクトルを帰属させる形式の分類手法であれば、非階層型クラスタリング以外の手法でも利用可能である。また、クラスタの重

心ベクトルと文書ベクトルとの類似度を測るための類似測度としては、余弦測度、内積測度、ユークリッド距離測度、マハラノビス距離測度などが利用可能である。

【0188】初期重心指定部2001によって、前記分類対象データと同一の特徴次元数をもつ任意の複数の文書ベクトルがクラスタリングの初期重心として入力される。前記任意の文書ベクトルは操作者により指定することもできるし、また分類対象の文書特徴ベクトルなどに基づいて構築した規則を操作者が選択することにより間接的に文書ベクトルを指定することもできる。

【0189】また、前記任意の文書ベクトルとしては、前記文書特徴ベクトルと同じ次元数を有するベクトル形式であれば、どのようなものでも適用可能である。また、任意の文書ベクトルとしては、そのほかに、より操作者にとって理解しやすいものとして、分類対象文書群以外の文書を文書特徴ベクトルと同じ次元をもつベクトルに表現したものをもちいることができる。

【0190】また、任意の文書ベクトルとしてそのほかには、文書特徴ベクトルをもちいることができる。任意の文書ベクトルとしてそのほかには、解析部402によって抽出される単語や単語品詞情報をもちいることができる。また、任意の文書ベクトルとしてそのほかには、分類結果記憶部407によって記憶されている事前におこなわれた分類結果である分類代表値をもちいることができる。

【0191】上記の指示情報は、それぞれ単独でもちいるほか、それらを適切に組み合わせたものをもちいるようにしてもよい。

【0192】2つの任意の文書ベクトル、3つの文書特徴ベクトル、一つの単語、一つの分類代表値とそれらの組み合わせ規則を指定することで、5つの初期クラスタ重心を求める例を図22に示す。図22に示すとおり、本実施の形態では、初期クラスタ重心1として文書1を、初期クラスタ重心2として文書2と文書3の平均を、初期クラスタ重心3として文書4と単語1の平均を、初期クラスタ重心4として文書5を、初期クラスタ重心5として分類代表値1を各々指定している。

【0193】また、指定された文書ベクトルが、操作者が指定したクラスタ数に満たない場合には、k－means法などでもちいられている一般的な自動初期重心選出法をもちいて残りのクラスタ重心を求めることができる。このようにして求めた初期重心に基づいてk－means法等をもちいて、クラスタの精練化をおこなうことで文書分類を実行する。

【0194】以上説明したように、この実施の形態6によれば、文書分類手法として、非階層型クラスタリング手法をもちいて、その際に必要となる初期クラスタ重心を、操作者が任意に指定することができ、その指定された初期クラスタ重心にしたがって文書分類をおこなうので、操作者の意図を反映する文書分類をおこなうことができる。

【0195】

【発明の効果】以上説明したように、請求項1の発明によれば、分類対象である文書群での文書間の類似性に基づいて、各文書をそれら文書間の意味的な関連性を反映しうる表現空間へ変換するための表現空間変換関数を算出し、その表現空間で文書分類をおこなうことにより、操作者の意図を反映しうる文書分類を実現することが可能な文書分類装置が得られるという効果を奏する。

【0196】また、請求項2の発明によれば、表現空間変換関数を導出する際に必要となる文書間の類似性として文書特徴ベクトル間の内積をもちいることにより、文書間の意味的な関連性を反映した文書分類をおこなうことが可能な文書分類装置が得られるという効果を奏する。

【0197】また、請求項3の発明によれば、表現空間変換関数を導出する際に必要となる文書間の類似性として文書特徴ベクトル間の内積に加え、文書の作成者や作成日などの文書間類似情報ももちいることにより、文書間の意味的な関連性を反映した文書分類をおこなうことが可能な文書分類装置が得られるという効果を奏する。

【0198】また、請求項4の発明によれば、算出する文書特徴ベクトルと表現空間変換関数を記憶することにより、表現空間変換関数を算出する部分と実際に前記表現空間変換関数をもちいて変換された文書をもちいて文書分類をおこなう部分とを分離して処理するので、その都度、表現空間変換関数を算出することなしに文書分類を実行でき、さらに、前記文書特徴ベクトル変換部でもちいる表現空間変換関数として、事前に他の文書特徴ベクトルに基づいて生成された表現空間変換関数をもちいることもできるため、文書分類の繰り返し実行を短時間で効率良くおこなうことが可能な文書分類装置が得られるという効果を奏する。

【0199】また、請求項5の発明によれば、文書分類の繰り返し実行をおこなう際、個々の分類実行ごとに、文書特徴ベクトルやそれらを構成する特徴次元を操作することで、各分類ごとに異なる単語を削除して文書分類を実行する等の分類対象文書の範囲の変更や分類をおこなう空間の変更をおこなうことが可能な文書分類装置が得られるという効果を奏する。

【0200】また、請求項6の発明によれば、表現空間変換関数が文書特徴ベクトルの内積をに基づいて算出される場合、表現空間変換関数をもちいて変換された文書をもちいて文書分類をおこなう部分において、文書特徴ベクトルやその特徴次元が操作された場合に生じる表現空間変換関数の不整合を簡便に修正することができるので、より適正な文書特徴ベクトルの変換をおこなうことが可能な文書分類装置が得られるという効果を奏する。

【0201】また、請求項7の発明によれば、前記表現空間変換関数をもちいて構成される空間の特徴次元につ

いて操作者が簡便な操作をすることにより、操作者の意図を反映しうる文書分類をおこなうことが可能な文書分類装置が得られるという効果を奏する。

【0202】また、請求項8の発明によれば、前記表現空間変換関数をもちいて構成される空間の特徴次元について、操作者により指示された分類対象以外の任意の文書ベクトルデータをもちいての簡便な操作をすることにより、操作者の意図を反映しうる文書分類をおこなうことが可能な文書分類装置が得られるという効果を奏する。

【0203】また、請求項9の発明によれば、前記表現空間変換関数をもちいて構成される空間の特徴次元について、操作者により指示された文書特徴ベクトルをもちいての簡便な操作をすることにより、操作者の意図を反映しうる文書分類をおこなうことが可能な文書分類装置が得られるという効果を奏する。

【0204】また、請求項10の発明によれば、前記表現空間変換関数をもちいて構成される空間の特徴次元を、操作者により指示された解析情報をもちいての簡便な操作をすることにより、操作者の意図を反映しうる文書分類をおこなうことが可能な文書分類装置が得られるという効果を奏する。

【0205】また、請求項11の発明によれば、前記表現空間変換関数をもちいて構成される空間の特徴次元を、操作者により指示された事前に分類された分類結果をもちいての簡便な操作をすることにより、操作者の意図を反映しうる文書分類をおこなうことが可能な文書分類装置が得られるという効果を奏する。

【0206】また、請求項12の発明によれば、文書分類手法として、非階層型クラスタリング手法をもちいて、その際に必要となる初期クラスタ重心を、操作者が任意に指定することができ、その指定された初期クラスタ重心にしたがって文書分類をおこなうので、操作者の意図を反映する文書分類をおこなうことが可能な文書分類装置が得られるという効果を奏する。

【0207】また、請求項13の発明によれば、文書分類手法として、非階層型クラスタリング手法をもちいて、その際に必要となる初期クラスタ重心として、分類対象以外の任意の文書をもちいることができるので、操作者の意図を反映する文書分類をおこなうことが可能な文書分類装置が得られるという効果を奏する。

【0208】また、請求項14の発明によれば、文書分類手法として、非階層型クラスタリング手法をもちいて、その際に必要となる初期クラスタ重心として、文書特徴ベクトルをもちいることができるので、操作者の意図を反映する文書分類をおこなうことが可能な文書分類装置が得られるという効果を奏する。

【0209】また、請求項15の発明によれば、文書分類手法として、非階層型クラスタリング手法をもちいて、その際に必要となる初期クラスタ重心として、分類対象文書を文書解析部に作用させた結果得られる単語等の解析情報をもちいることができるので、操作者の意図を反映する文書分類をおこなうことが可能な文書分類装置が得られるという効果を奏する。

【0210】また、請求項16の発明によれば、文書分類手法として、非階層型クラスタリング手法をもちいて、その際に必要となる初期クラスタ重心として、事前に分類された分類結果をもちいることができるので、操作者の意図を反映する文書分類をおこなうことが可能な文書分類方法が得られるという効果を奏する。

【0211】また、請求項17の発明によれば、分類対象である文書群での文書間の類似性に基づいて、各文書をそれら文書間の意味的な関連性を反映しうる表現空間へ変換するための表現空間変換関数を算出し、その表現空間で文書分類をおこなうことにより、操作者の意図を反映しうる文書分類を実現することが可能な文書分類方法が得られるという効果を奏する。

【0212】また、請求項18の発明によれば、表現空間変換関数を導出する際に必要となる文書間の類似性として文書特徴ベクトル間の内積をもちいることにより、文書間の意味的な関連性を反映した文書分類をおこなうことが可能な文書分類方法が得られるという効果を奏する。

【0213】また、請求項19の発明によれば、表現空間変換関数を導出する際に必要となる文書間の類似性として文書特徴ベクトル間の内積に加え、文書の作成者や作成日などの文書間類似情報ももちいることにより、文書間の意味的な関連性を反映した文書分類をおこなうことが可能な文書分類方法が得られるという効果を奏する。

【0214】また、請求項20の発明によれば、算出する文書特徴ベクトルと表現空間変換関数を記憶することにより、表現空間変換関数を算出する部分と実際に前記表現空間変換関数をもちいて変換された文書をもちいて文書分類をおこなう部分とを分離して処理するので、その都度、表現空間変換関数を算出することなしに文書分類を実行でき、さらに、前記文書特徴ベクトル変換部でもちいる表現空間変換関数として、事前に他の文書特徴ベクトルに基づいて生成された表現空間変換関数をもちいることもできるため、文書分類の繰り返し実行を短時間で効率良くおこなうことが可能な文書分類方法が得られるという効果を奏する。

【0215】また、請求項21の発明によれば、文書分類の繰り返し実行をおこなう際、個々の分類実行ごとに、文書特徴ベクトルやそれらを構成する特徴次元を操作することで、各分類ごとに異なる単語を削除して文書分類を実行する等の分類対象文書の範囲の変更や分類をおこなう空間の変更をおこなうことが可能な文書分類方法が得られるという効果を奏する。

【0216】また、請求項22の発明によれば、表現空

間変換関数が文書特徴ベクトルの内積をに基づいて算出される場合、表現空間変換関数をもちいて変換された文書をもちいて文書分類をおこなう部分において、文書特徴ベクトルやその特徴次元が操作された場合に生じる表現空間変換関数の不整合を簡便に修正することができるので、より適正な文書特徴ベクトルの変換をおこなうことが可能な文書分類方法が得られるという効果を奏する。

【０２１７】また、請求項２３の発明によれば、前記表現空間変換関数をもちいて構成される空間の特徴次元について操作者が簡便な操作をすることにより、操作者の意図を反映しうる文書分類をおこなうことが可能な文書分類方法が得られるという効果を奏する。

【０２１８】また、請求項２４の発明によれば、前記表現空間変換関数をもちいて構成される空間の特徴次元について、操作者により指示された分類対象以外の任意の文書ベクトルデータをもちいての簡便な操作をすることにより、操作者の意図を反映しうる文書分類をおこなうことが可能な文書分類方法が得られるという効果を奏する。

【０２１９】また、請求項２５の発明によれば、前記表現空間変換関数をもちいて構成される空間の特徴次元について、操作者により指示された文書特徴ベクトルをもちいての簡便な操作をすることにより、操作者の意図を反映しうる文書分類をおこなうことが可能な文書分類方法が得られるという効果を奏する。

【０２２０】また、請求項２６の発明によれば、前記表現空間変換関数をもちいて構成される空間の特徴次元を、操作者により指示された解析情報をもちいての簡便な操作をすることにより、操作者の意図を反映しうる文書分類をおこなうことが可能な文書分類方法が得られるという効果を奏する。

【０２２１】また、請求項２７の発明によれば、前記表現空間変換関数をもちいて構成される空間の特徴次元を、操作者により指示された事前に分類された分類結果をもちいての簡便な操作をすることにより、操作者の意図を反映しうる文書分類をおこなうことが可能な文書分類方法が得られるという効果を奏する。

【０２２２】また、請求項２８の発明によれば、文書分類手法として、非階層型クラスタリング手法をもちいて、その際に必要となる初期クラスタ重心を、操作者が任意に指定することができ、その指定された初期クラスタ重心にしたがって文書分類をおこなうので、操作者の意図を反映する文書分類をおこなうことが可能な文書分類方法が得られるという効果を奏する。

【０２２３】また、請求項２９の発明によれば、文書分類手法として、非階層型クラスタリング手法をもちいて、その際に必要となる初期クラスタ重心として、分類対象以外の任意の文書をもちいることができるので、操作者の意図を反映する文書分類をおこなうことが可能な文書分類方法が得られるという効果を奏する。

【０２２４】また、請求項３０の発明によれば、文書分類手法として、非階層型クラスタリング手法をもちいて、その際に必要となる初期クラスタ重心として、文書特徴ベクトルをもちいることができるので、操作者の意図を反映する文書分類をおこなうことが可能な文書分類方法が得られるという効果を奏する。

【０２２５】また、請求項３１の発明によれば、文書分類手法として、非階層型クラスタリング手法をもちいて、その際に必要となる初期クラスタ重心として、分類対象文書を文書解析部に作用させた結果得られる単語等の解析情報をもちいることができるので、操作者の意図を反映する文書分類をおこなうことが可能な文書分類方法が得られるという効果を奏する。

【０２２６】また、請求項３２の発明によれば、文書分類手法として、非階層型クラスタリング手法をもちいて、その際に必要となる初期クラスタ重心として、事前に分類された分類結果をもちいることができるので、操作者の意図を反映する文書分類をおこなうことが可能な文書分類方法が得られるという効果を奏する。

【０２２７】また、請求項３３の発明によれば、請求項１７～３２に記載された方法をコンピュータに実行させるプログラムを記録したことで、そのプログラムを機械読み取り可能となり、これによって、請求項１７～３２の動作をコンピュータによって実現することが可能な記録媒体が得られるという効果を奏する。

【図面の簡単な説明】

【図１】この発明の実施の形態１による文書分類装置を構成する情報処理システム全体のハードウエア構成を示す説明図である。

【図２】実施の形態１による文書分類装置を構成する情報処理システムにおけるサーバーをハードウエア的に示す説明図である。

【図３】実施の形態１による文書分類装置を構成する情報処理システムにおけるクライアントをハードウエア的に示す説明図である。

【図４】実施の形態１による文書分類装置の構成を機能的に示すブロック図である。

【図５】実施の形態１による文書分類装置の構成を機能的に示す別のブロック図である。

【図６】実施の形態１による文書分類装置の構成を機能的に示す別のブロック図である。

【図７】実施の形態１による文書分類装置の文書－単語行列データと文書特徴ベクトルの一例を示す説明図である。

【図８】実施の形態１による文書分類装置の一連の処理の手順を示すフローチャートである。

【図９】実施の形態１による文書分類装置の一連の処理の別の手順を示すフローチャートである。

【図１０】この発明の実施の形態２による文書分類装置

35

の構成を機能的に示すブロック図である。

【図１１】実施の形態２による文書分類装置の一連の処理の手順を示すフローチャートである。

【図１２】この発明の実施の形態３による文書分類装置の構成を機能的に示すブロック図である。

【図１３】実施の形態３による文書分類装置のベクトル修正部の処理内容の手順を示すフローチャートである。

【図１４】実施の形態３による文書分類装置の文書特徴ベクトルから特徴次元を削除する手続きの一例を示す説明図である。

【図１５】実施の形態３による文書分類装置の一連の処理の手順を示すフローチャートである。

【図１６】この発明の実施の形態４による文書分類装置の構成を機能的に示すブロック図である。

【図１７】実施の形態４による文書分類装置の一連の処理の手順を示すフローチャートである。

【図１８】この発明の実施の形態５による文書分類装置の構成を機能的に示すブロック図である。

【図１９】実施の形態５による文書分類装置の一連の処理の一部の手順を示すフローチャートである。

【図２０】この発明の実施の形態６による文書分類装置の構成を機能的に示すブロック図である。

【図２１】実施の形態６による文書分類装置の一連の処理の一部の手順を示すフローチャートである。

【図２２】実施の形態６による文書分類装置の初期クラスタ重心を求める処理の一例についての説明図である。

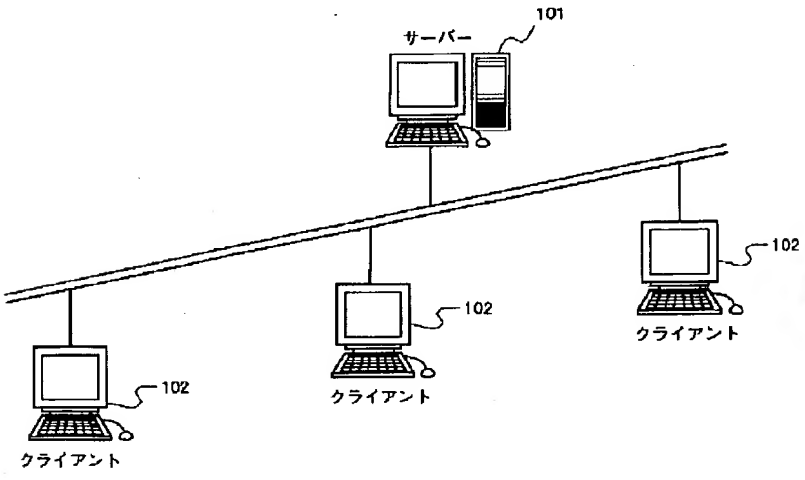
【符号の説明】

１０１　サーバー

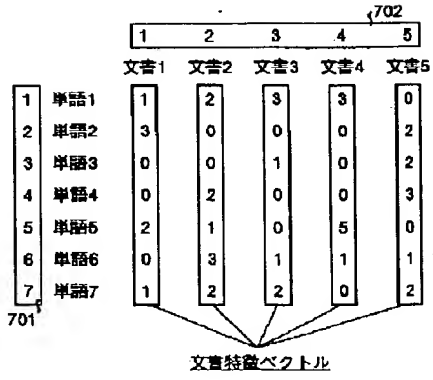
36

１０２　クライアント
１０３　ネットワーク
２０１　ＣＰＵ
２０４　Ｉ／Ｆ
２０６　ディスク装置
３０１　ＣＰＵ
３０６　ハードディスク
３０８　ディスプレイ
３０９　Ｉ／Ｆ
３１１　キーボード
３１２　マウス
３１３　スキャナ
４０１　入力部
４０２　解析部
４０３　ベクトル生成部
４０４　変換関数算出部
４０５　ベクトル変換部
４０６　分類部
４０７　分類結果記憶部
４２１　内積算出部
４３１　文書間類似情報設定部
１００１　ベクトル記憶部
１００２　変換関数記憶部
１２０１　ベクトル修正部
１６０１　変換関数修正部
１８０１　変換関数修正指示部
２００１　初期重心指定部
２００２　初期重心登録部

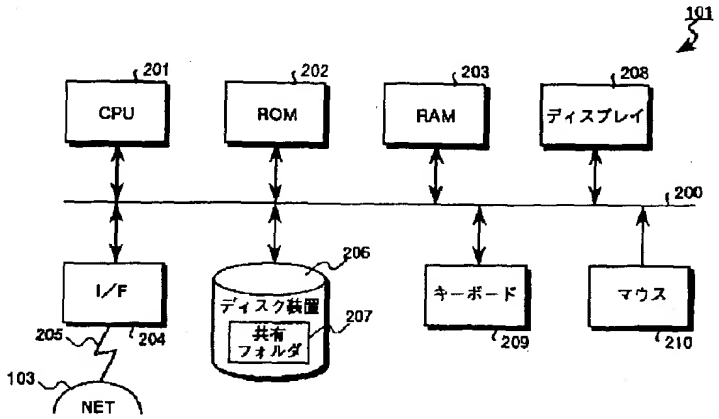
【図１】



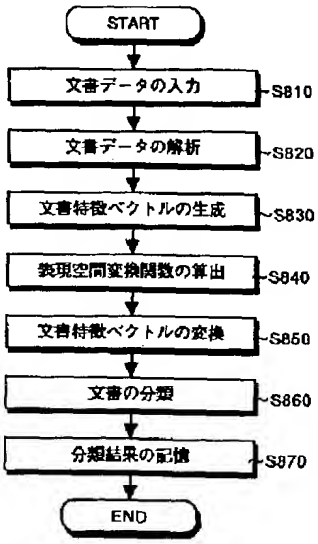
【図７】



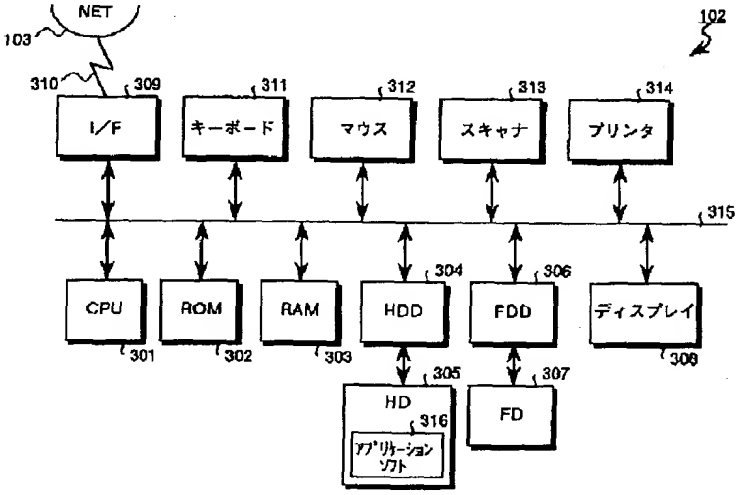
【図２】


101
CPU
201
ROM
202
RAM
203
ディスプレイ
208
200
I/F
204
205
103
NET
ディスク装置
206
共有フォルダ
207
キーボード
209
マウス
210

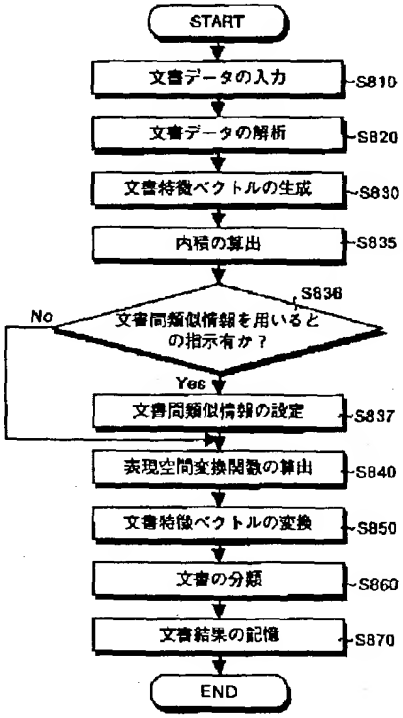
【図８】


START
文書データの入力
S810
文書データの解析
S820
文書特徴ベクトルの生成
S830
表現空間変換関数の算出
S840
文書特徴ベクトルの変換
S850
文書の分類
S860
分類結果の記憶
S870
END

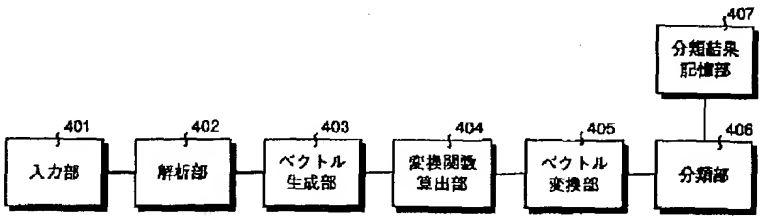
【図３】


NET
103
102
310
I/F
309
キーボード
311
マウス
312
スキャナ
313
プリンタ
314
315
CPU
301
ROM
302
RAM
303
HDD
304
FDD
306
ディスプレイ
308
HD
305
316
アプリケーションソフト
FD
307

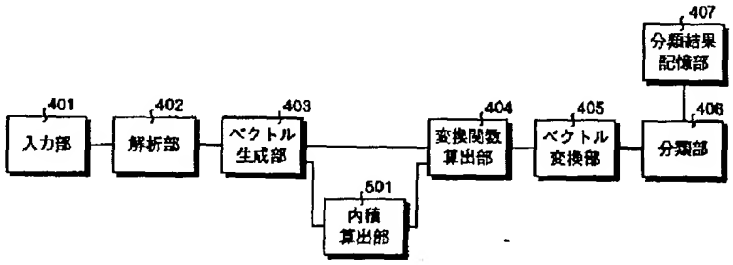
【図９】


START
文書データの入力
S810
文書データの解析
S820
文書特徴ベクトルの生成
S830
内積の算出
S835
S836
No
文書間類似情報を用いるとの指示有か？
Yes
文書間類似情報の設定
S837
表現空間変換関数の算出
S840
文書特徴ベクトルの変換
S850
文書の分類
S860
文書結果の記憶
S870
END

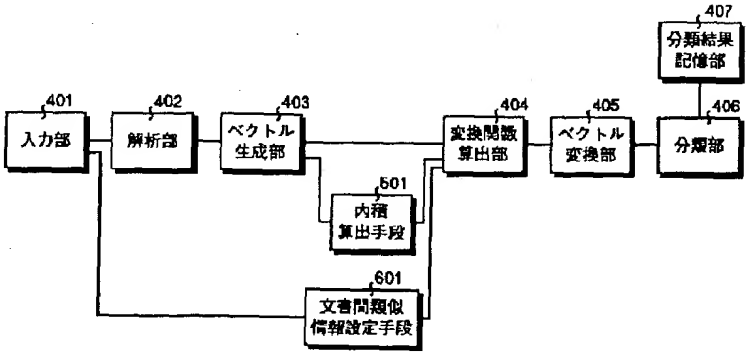
【図４】


407
分類結果記憶部
401
入力部
402
解析部
403
ベクトル生成部
404
変換関数算出部
405
ベクトル変換部
406
分類部

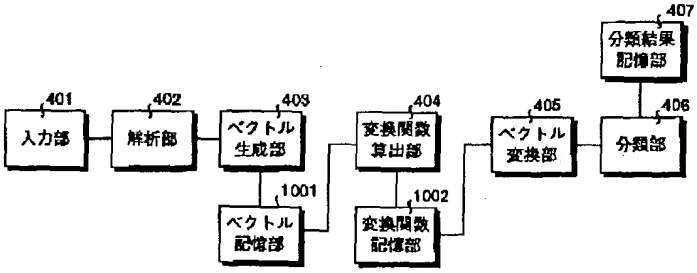
【図５】



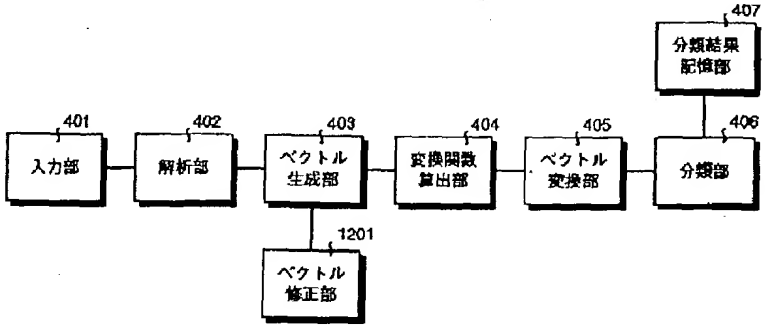
【図６】



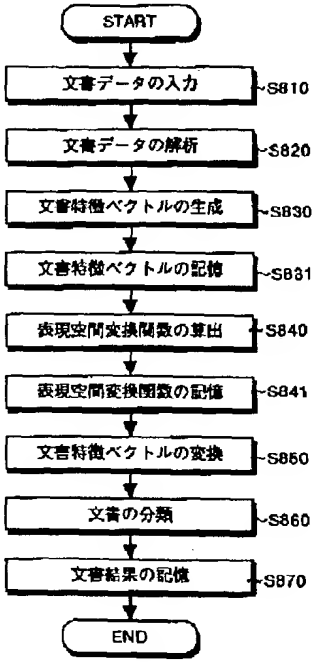
【図１０】



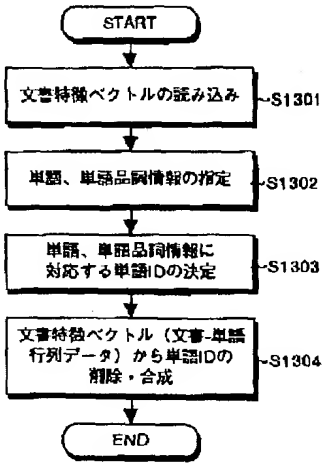
【図１２】



【図１１】



【図１３】



【図１４】

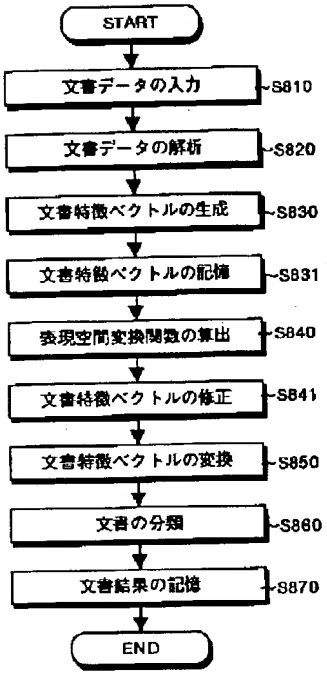
単語2と単語5を削除

		文書1	文書2	文書3	文書4	文書5
1	単語1	1	2	3	3	0
2	単語2	3	0	0	0	2
3	単語3	0	0	1	0	2
4	単語4	0	2	0	0	3
5	単語5	2	1	0	5	0
6	単語6	0	3	1	1	1
7	単語7	1	2	2	0	2

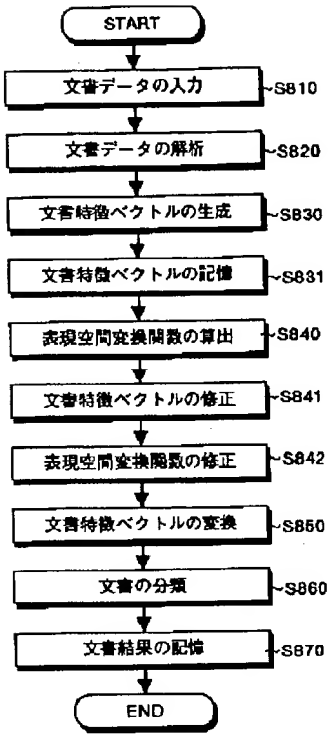
↓

		文書1	文書2	文書3	文書4	文書5
1	単語1	1	2	3	3	0
3	単語3	0	0	1	0	2
4	単語4	0	2	0	0	3
6	単語6	0	3	1	1	1
7	単語7	1	2	2	0	2

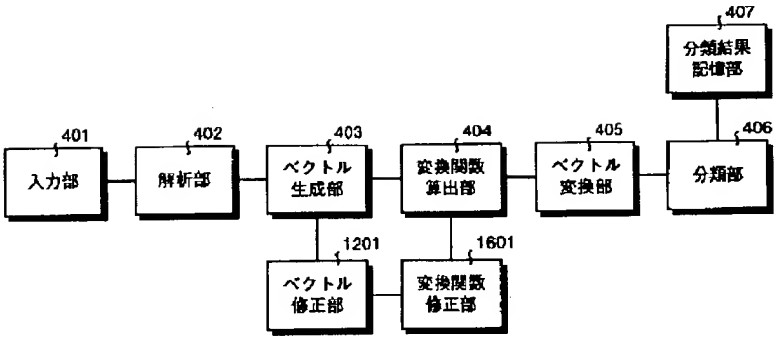
【図１５】



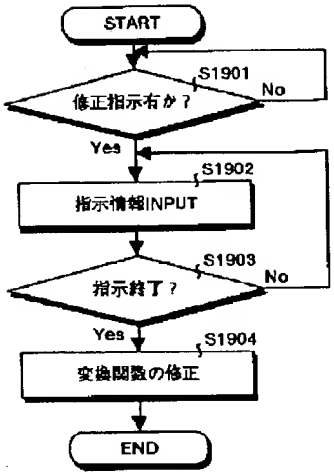
【図１７】



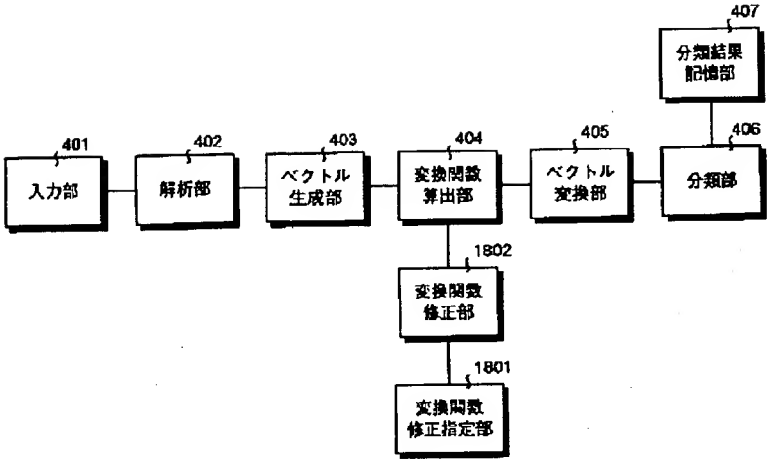
【図１６】



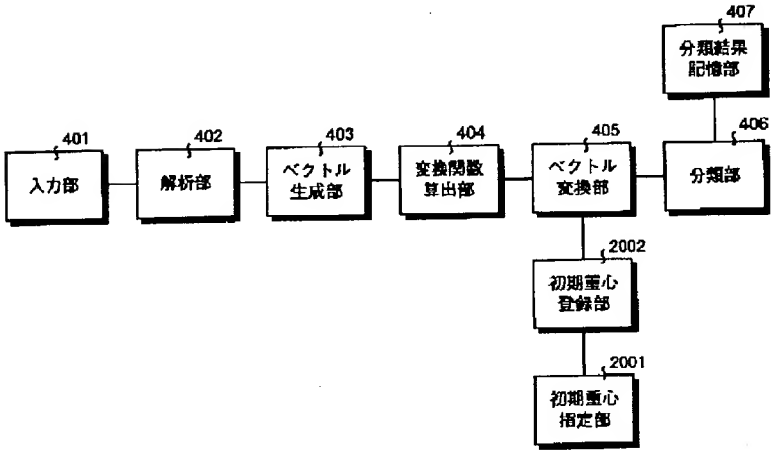
【図１９】



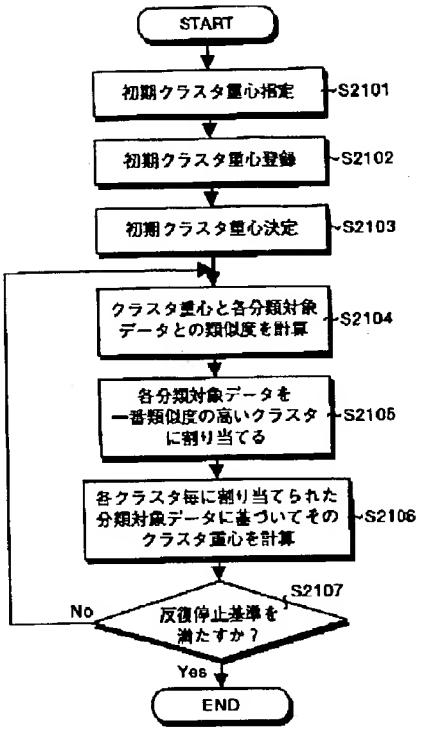
【図１８】



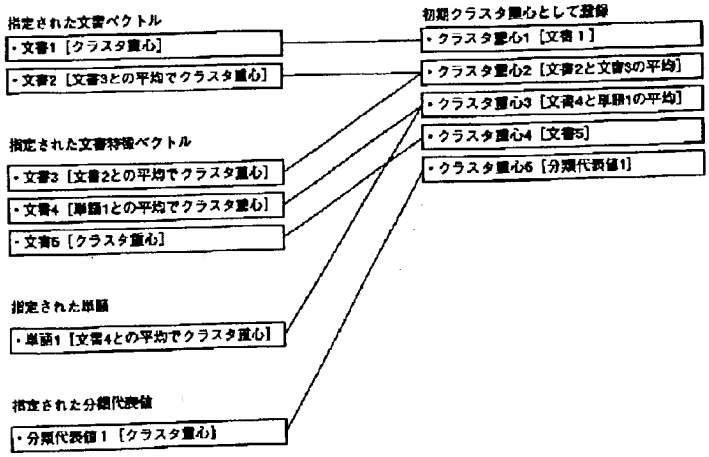
【図２０】



【図２１】



【図２２】



フロントページの続き

(72)発明者　武谷　一寿
東京都大田区中馬込１丁目３番６号　株式会社リコー内
(72)発明者　中島　明子
東京都大田区中馬込１丁目３番６号　株式会社リコー内
(72)発明者　長束　哲郎
東京都大田区中馬込１丁目３番６号　株式会社リコー内
(72)発明者　山崎　真湖人
東京都大田区中馬込１丁目３番６号　株式会社リコー内
(72)発明者　藤田　克彦
東京都大田区中馬込１丁目３番６号　株式会社リコー内